Introduction to Biostatistics: Part 1, Basic Concepts

Statistical methods commonly used to analyze data presented in journal articles should be understood by both medical scientists and practicing clinicians. Inappropriate data analysis methods have been reported in 42% to 78% of original publications in critical reviews of selected medical journals. The only way to halt researchers' misuse of statistics and improve the clinician's knowledge of statistics is through education. This is the first of a six-part series of articles intended to provide the reader with a basic, yet fundamental knowledge of common biomedical statistical methods. The series will cover basic concepts of statistical analysis, descriptive statistics, statistical inference theory, comparison of means, χ^2, and correlational and regression techniques. A conceptual explanation will accompany discussion of the appropriate use of these techniques. [Gaddis ML, Gaddis GM: Introduction to biostatistics: Part 1, basic concepts. Ann Emerg Med January 1990;19:86-89.]

Monica L Gaddis, PhD
Gary M Gaddis, MD, PhD
Kansas City, Missouri

From the Departments of Surgery and Emergency Health Services, Truman Medical Center, University of Missouri, Kansas City.

Received for publication September 1, 1989. Accepted for publication October 4, 1989.

Address for reprints: Monica L Gaddis, PhD, Department of Surgery, Truman Medical Center, 2301 Holmes, Kansas City, Missouri 64108.

INTRODUCTION

The practicing physician must remain abreast of new information and techniques and meet standards for continuing medical education. A major source of new information is the medical journal. Given the competition involved in achieving publication of research, as well as the review and editing process, the reader might justifiably assume that published studies are correct in their methodology, regardless of the conclusions made. Although this assumption seems reasonable, it is not correct. Between 1979 and 1984, 42% to 78% of original publications from selected medical journals used inappropriate statistical analysis methods.[1-5]

Statistical analysis involves organization and mathematical manipulation of data. This process can describe characteristics studied and help to infer conclusions from the data, thus guiding the acceptance or rejection of a given treatment or theory. Incorrect data analysis is a grave error in the research process, often leading to inappropriate conclusions, continued study of erroneous hypotheses, and curtailed study of viable therapies and therapeutic adjuncts. Additional dangers ensue when a physician uses nonefficacious treatment on a patient. From this, it becomes apparent that a basic knowledge of statistics can be an important tool for any clinician, whether in performing research or simply reading about it.

However, statistical analysis of data is a task commonly delegated to statistical consultants. This is often justified by citing that there are those more qualified to perform this function than the principal investigator of a study. Though this seems logical, the principal investigator remains the individual ultimately responsible for the content and conclusions of a research project. The principal investigator cannot effectively meet this obligation fully if he does not have an adequate working knowledge of biomedical statistics. The investigator cannot plead innocence through ignorance when serious errors are made.

Thus, there is an obvious need for statistical education among clinicians, not only to provide for a better understanding when reading the biomedical literature but also to aid medical researchers in communication with consulting statisticians and for selection of appropriate data analysis techniques.

FIGURE 1. *A bell-shaped or normal distribution curve.*

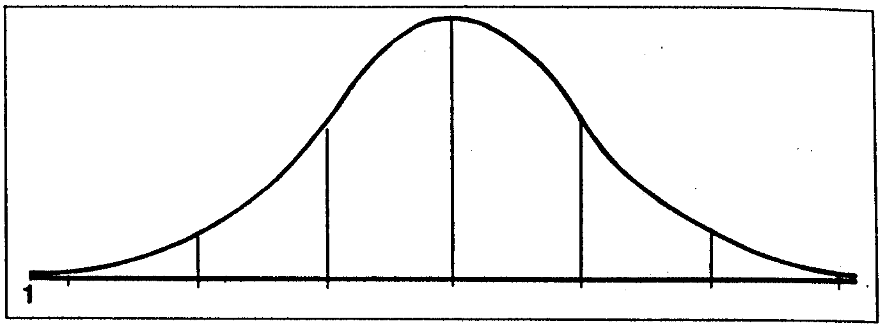
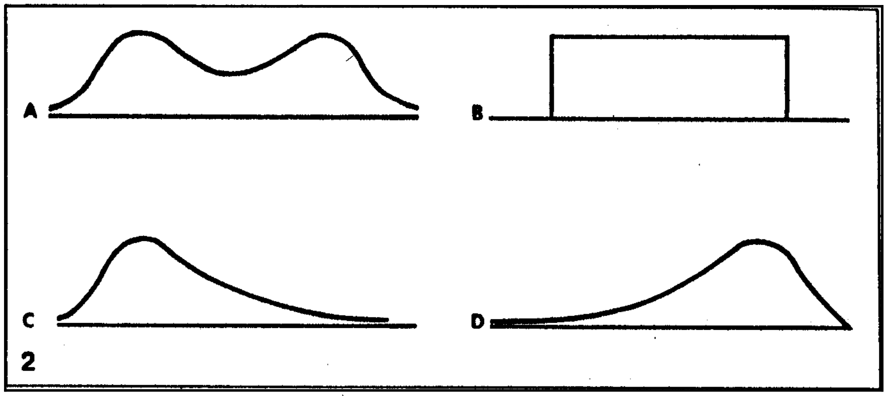


FIGURE 2. *Frequency distributions:* A, *bimodal;* B, *rectangular;* C, *positively skewed;* D, *negatively skewed. (Hopkins KD, Glass GV:* Basic Statistics for the Behavioral Sciences. *Englewood Cliffs, New Jersey, Prentice-Hall Inc, 1978, p 36.)*



STATEMENT OF PURPOSE

It is our purpose to present a six-part series discussing the basics of biomedical statistics, with the intent to familiarize the reader with the terminology and appropriate use of data analysis techniques commonly used in original papers published in *Annals of Emergency Medicine* and other clinical medical journals. Learning will be directed toward a conceptual understanding of statistical analysis methods rather than computational exercises.

Part 1, in this issue of *Annals,* presents some basic concepts of statistical analysis. Knowledge of these building blocks is necessary for a complete understanding of biomedical statistics and the topics of future articles in this series.

Part 2 will address descriptive statistics. These include measures of central tendency (mean, median, and mode), measures of variability (standard deviation and standard error of the mean), and confidence intervals. Appropriate uses of these will be discussed.

Part 3 will introduce statistical inference theory. An explanation of the concept of hypothesis testing, definition of the probability value *P,* a discussion of the terms alpha, beta, and power, and a discussion of clinical versus statistical significance will be included. Sensitivity, specificity, and predictive value will also be addressed.

Part 4 will present parametric and nonparametric methods used for the comparison of means. Included will be analysis of variance (ANOVA), the Student's *t* test, the Mann-Whitney *U* test, and methods of multiple comparisons.

Part 5 will present a discussion of χ^2 and Fisher's exact tests. Both statistical methods should be understood by readers of biomedical research involving a study of the efficacy of experimental medical treatments.

Part 6 will give a basic discussion of correlation and regression, along with the series' concluding remarks.

SAMPLE VERSUS POPULATION

It is clearly impossible for most scientific studies to survey all individuals of a group about which conclusions are to be drawn. Cost and time considerations force the researcher to settle for studying a subset of a group in order to form conclusions about the entire group. For example, if one wished to know the systolic, diastolic, and mean blood pressures of the entire population of the United States, this testing could be done as part of the next census, in which all members of the population would be surveyed. However, the cost of training census takers for this time-consuming task and the need to hire additional census takers because of the increased time required to check each individual would make sampling the entire population an impractical task. Similarly, in biomedical research, time and financial costs preclude study of every member of the population.

However, if a representative sample of an appropriate population can be obtained and studied, conclusions about the sample can be properly extrapolated to the defined population. The key to obtaining a representative sample of that population lies in random selection of a study sample from the applicable population. Randomization can be accomplished by sample subject selection using a random number table, drawing numbers or names out of a hat, or the like. Random sampling implies that all individuals in a population have an equal chance to be included in the sample.

It is when random allocation of study subjects to treatment groups is violated that bias is introduced. Bias can easily lead to erroneous conclusions because control and treatment groups may inherently differ in relevant characteristics before the study is initiated. Therefore, post-treatment differences between groups can erroneously be ascribed to an effect of the experimental treatment! Improper allocation of subjects to control and treatment groups remains a significant problem confounding current biomedical research.

DATA SCALES

Before an analysis method can be selected, the type of data that will be generated by the research process must be defined. Data will fit a nom-

Statistical analysis The organization and mathematical manipulation of data, used to describe characteristics studied and/or to help infer conclusions from the data

Population A large group possessing a given characteristic or set of characteristics. A population may be finite (the states of the United States) or infinite (blood pressure measurements of all infants born in New York)[6]

Sample The studied subset of members of a defined population[6]

Random sample The process of selection of a sample from a population whereby each member of the population has an equal and independent chance of being chosen[6]

Nominal scale Numbers are arbitrarily assigned to characteristics for data classification

Ordinal scale Numbers are used to denote rank-order, without defining a magnitude of difference between numbers

Interval scale Numbers denote units of equal magnitude as well as rank order on a scale *without* an absolute zero

Ratio scale Numbers denote units of equal magnitude as well as rank order on a scale *with* an absolute zero

Distribution The systematic organization of a collection of data

Normal distribution A grouping of data that is graphically symmetrical and bell shaped. Many human anatomic and physiologic characteristics are normally distributed

Parametric methods Used when the data studied are from a sample or population that is normally distributed. Data must be of an interval or ratio scale

Nonparametric methods Used when the data studied are from a sample or population that deviates from a normal distribution. Ordinal data are analyzed using nonparametric methods of analysis

3

FIGURE 3. *Summary of biostatistical terms.*

inal, ordinal, interval, or ratio scale.

Nominal Scale

Nominal is the most primitive of the data scales. Information classified by an assigned number or code to make the data numeric fits a nominal scale. For instance, gender may be defined as 1 for "female" and 2 for "male." Clinical diagnoses may also be assigned representative numbers, such as 1 for "renal failure," 2 for "congestive heart failure," 3 for "diabetes," and so forth. It must be emphasized that the numbers selected are purely arbitrary and are chosen at the discretion of the researcher without regard to any order of ranking of severity.[6,7]

Ordinal Scale

Ordinal scale data can be ranked in a specific order, be it low to high or high to low. An example would be data from a questionnaire in which a response of "strongly agree" is scored as 5, "agree" is scored 4, "no opinion" is scored 3, "disagree" is scored 2, and "strongly disagree" is scored 1. In this example, the responses are scored on a continuum, without a consistent level of magnitude of differences between ranks. However, unlike the nominal scale, numbers of an ordinal scale are not arbitrary because the order of numbers is meaningful.[6] For instance, in the above example, progressively larger numbers indicate progressively greater agreement with the question presented. The Glasgow Coma Score[8] allots a progressively decreased classification number that implies progressively worsened obtundation. Other examples of ordinal scales familiar to many emergency physicians would include the Trauma Score[9] and the Injury Severity Score.[10]

Average values of ordinal scale data are often calculated but are usually misleading because of the lack of any consistent magnitude of difference between units of the scale.[11] Also, it is not uncommon in trauma outcome studies to see such data reported with standard deviation values. Calculation of such statistics from nonparametric, and thus nonnormally distributed data, is highly questionable because use of the standard deviation assumes that the data are normally distributed.[11]

Interval Scale

Interval scale data are a step more sophisticated than ordinal scale data. Not only is there a predetermined order to the numbering of the scale, but also there is a consistent level of magnitude of difference described between the observed data units.[6] Interval data also have a clearly defined unit of measure. However, "the zero point on the scale is arbitrary, and does not correspond to a total absence of the characteristic measured . . . ."[6] The Farenheit scale for temperature is interval in nature, as the numbering of the scale is consistent, yet the zero value is arbitrary. Because of the consistent numbering of the scale and equal magnitude between measurement units, average values and measures of variability of interval scale data are meaningful. An interval scale can be converted to an ordinal scale to show ranking, but an ordinal scale ordinarily cannot be converted to an interval scale.[6]

Ratio Scale

The ratio scale is simply an interval scale with an absolute zero.[6] A predetermined order to the numbering of the scale is present, as is a consistent level of magnitude between each unit of measure. Ratio data can be converted to ordinal data. Heart rate, blood pressure, distance, time, and degrees Kelvin represent examples of ratio scale data.

DISTRIBUTIONS

Once data are collected, they can be organized into a distribution, or graph of frequency of occurrence. This is a visually descriptive tool that allows the researcher to begin to define and analyze data.

Figure 1 is a theoretical frequency distribution of resting heart rate of emergency physicians. Its shape is symmetrical and bell-shaped, and is defined as the "normal distribution." Many human anatomic and physiologic characteristics approach the normal distribution.[6] Other terminology used synonymously with "normal distribution" includes Gaussian curve, curve of error, and normal probability curve. An understanding of the characteristics of the normal distribution is fundamental in the development of even a basic knowledge of biostatistics. This topic will be discussed in greater detail in Part 2 of this series.

Other distributions are depicted in Figure 2.[6] Bimodal distributions have two peaks of cluster, or areas with a high frequency level. For example, if the weights of American adults were plotted, there would be two definitive points of cluster, one for female weight and one for male weight.

Data that are rectangularly distributed show equal frequency of occurrence for all levels of a characteristic. The date of birth (month and day) of a sample of all patients seen in an emergency department approaches this distribution pattern.

Skewed data are those that tail off to either the high or low end of measurement units. The annual income of patients seen in the ED of an inner-city community hospital will be defined by a distribution that is positively skewed, showing a high frequency of lower annual incomes. A negatively skewed distribution has a cluster of data on the high end of the unit scale and tails off toward the low end.[6]

Given the nature of data collected in medical science research, the normal distribution is the one with which the clinician will be most familiar. Most statistical methods applied to interval or ratio data assume that the data are normally distributed. It is when this assumption is violated that significant controversy exists in the statistical community regarding the proper application of statistical tests.

PARAMETRIC VERSUS NONPARAMETRIC METHODS

Statistical methods are defined as being parametric or nonparametric. The type of analysis method that should be selected is dependent on the nature of the data to be analyzed. Parametric methods use data extrapolated from a sample of the population studied to numerically describe some characteristic of a population. Parametric methods are valid only when that characteristic follows or nearly follows the normal distribution in the population studied.[12] Thus, parametric methods can be applied properly to most interval and ratio scale data when those data come from a sample of a normally distributed population. Parametric methods can be used to derive measures of central tendency and variability, which will be discussed in Part 2 of the series.

Nonparametric methods are applied to non-normally distributed data and/or data that do not meet the criteria for the use of parametric methods. Data that fit the ordinal scale definition should be analyzed by nonparametric methods. Examples include Glasgow Coma Scale,[8] Trauma Score,[9] the Injury Severity Score,[10] and other similar ordinal scales data.

SUMMARY

This has been the introductory installment of a series of articles outlining the proper use of biostatistics. We have introduced some basic concepts regarding data and its classification, which will be needed for understanding of topics to be presented subsequently (Figure 3). Measures of central tendency, measures of variability, confidence intervals, and the appropriate use of these statistical concepts will be discussed in part 2 of this series.

REFERENCES

1. Glantz SA: Biostatistics: How to detect, correct, and prevent errors in the medical literature. *Circulation* 1980;61:1-7.

2. Felson DT, Cupples LA, Meenan RF: Misuse of statistical methods in arthritis and rheumatism: 1982 versus 1967-68. *Arthritis Rheum* 1984;27:1018-1022.

3. Thorn MD, Pulliam CC, Symons MJ, et al: Statistical and research quality of the medical and pharmacy literature. *Am J Hosp Pharm* 1985;42:1077-1082.

4. Avram MJ, Shanks CA, Dykes MHM, et al: Statistical methods in anesthesia articles: An evaluation of two American journals during two six-month periods. *Anesth Analg* 1985;64:607-611.

5. MacArthur RD, Jackson GG: An evaluation of the use of statistical methodology in the *Journal of Infectious Diseases. J Infect Dis* 1984;149:349-354.

6. Hopkins KD, Glass GV: *Basic Statistics for the Behavioral Sciences.* Englewood Cliffs, New Jersey, Prentice-Hall Inc, 1978.

7. Elenbaas RM, Elenbaas JK, Cuddy PG: Evaluating the medical literature. Part II: Statistical analysis. *Ann Emerg Med* 1983;12:610-620.

8. Teasdale G, Jennett B: Assessment of coma and impaired consciousness: A practical scale. *Lancet* 1974;2:81-84.

9. Champion HR, Sacco WJ, Carnazzo AJ, et al: Trauma score. *Crit Care Med* 1981;9:672-676.

10. Baker SP, O'Neill B, Haddon W Jr, et al: The Injury Severity Score: A method for describing patients with multiple injuries and evaluating emergency care. *J Trauma* 1974;14:187-196.

11. Sokal RR, Rohlf FJ: *Biometry,* ed 2. New York, WH Freeman and Co, 1981, pp 39-52.

12. Glantz SA: *Primer of Biostatistics,* ed 2. New York, McGraw-Hill Book Co, 1987.

Introduction to Biostatistics: Part 2, Descriptive Statistics

Descriptive statistics include measures of central tendency and variability. Measures of central tendency include mean, median, and mode. The mean is the arithmetic average of data from interval or ratio scales. The median reflects the 50th percentile score. The mode is the most frequently occurring value of a data distribution. Measures of variability include range, interquartile range, standard deviation, and standard error of the mean. The range describes the spread between the extreme values of data. Interquartile range is data included between the 25th and 75th percentile of a distribution. Standard deviation describes variability of data about the sample mean, while standard error of the mean helps describe the distribution of several sample means about a true population mean. Finally, confidence intervals, which are derived from the standard error of the mean, define an interval likely to include a true population value, based on sample statistical values and probability characteristics of data distributions. [Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 2, descriptive statistics. Ann Emerg Med March 1990;19:309-315.]

Gary M Gaddis, MD, PhD
Monica L Gaddis, PhD
Kansas City, Missouri

From the Departments of Emergency Health Services and Surgery, Truman Medical Center, University of Missouri, Kansas City.

Received for publication September 1, 1989. Accepted for publication December 4, 1989.

Address for reprints: Monica L Gaddis, PhD, Department of Surgery, Truman Medical Center, 2301 Holmes, Kansas City, Missouri 64108.

INTRODUCTION

Statistical analysis is the process by which numerical data obtained by scientific inquiry are transformed into a useable form for scientific interpretation. This involves manipulation of data for describing characteristics studied (descriptive statistics) and transformation of the data to help infer conclusions from the data (inferential statistics).

This second of a six-part series on biostatistics focuses on descriptive statistics. A thorough understanding of this topic is needed before advancing to discussions about inferential statistics. Familiarity with the concepts regarding types of data and data distributions, as presented in part 1,[1] is required for understanding the concepts presented herein. Numerical examples are provided to facilitate understanding. Finally, there exist many common, yet inappropriate uses of statistics, which will be discussed in this article.

MEASURES OF CENTRAL TENDENCY

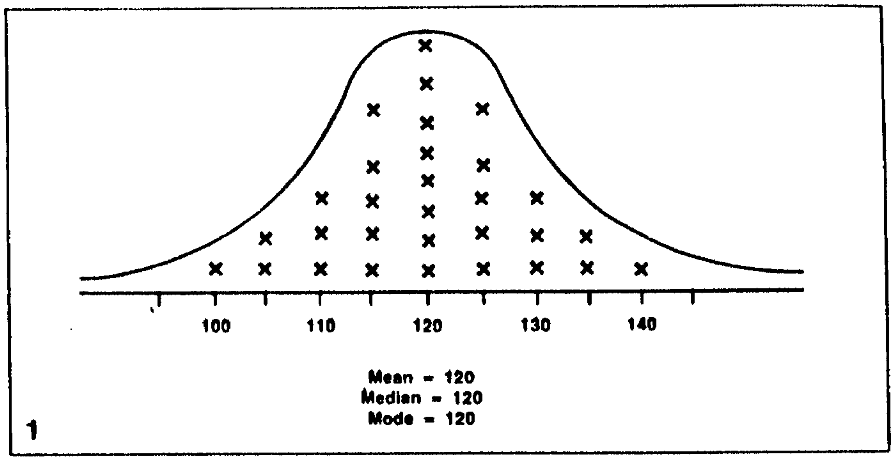
Mean

The mean is the arithmetic average of data and is expressed by the equation:

$$\bar{X} = \Sigma X_i/n$$

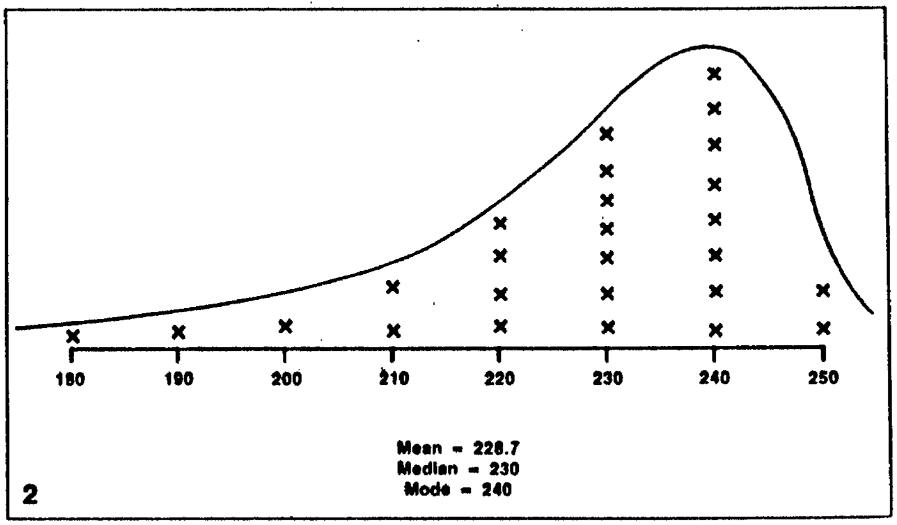
where $\bar{X}$ equals the mean, X_i equals each individual data point, and n is the number of data points in the sample. The mean can be calculated for interval and ratio scale data. However, mean values for ordinal scale data are generally misleading or invalid due to the lack of a consistent level of magnitude between numeric units of the scale.[1] Therefore, the mean is useful for data such as heart rate and blood pressure, but is misleading for arbitrarily constructed data scales such as the Apgar Scale, Glasgow Coma Score, and Trauma Score.[2,3]

The mean is affected by outliers, which are extreme values of a data distribution.[2] This is not true of other measures of central tendency.



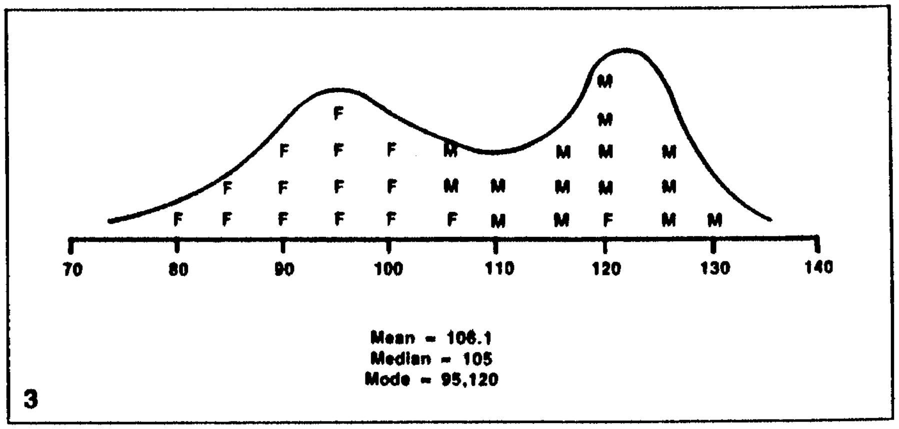
1

FIGURE 1. *Systolic blood pressure of 30 men aged 31 to 40 years. The mean, median, and mode all equal 120.*



2

FIGURE 2. *Systolic blood pressure in persons with renovascular hypertension. Mean, 228.7; median, 230; mode, 240.*



3

FIGURE 3. *Systolic blood pressure in young men and pregnant women. M denotes male subjects, F denotes female subjects. Mean, 106.1; median, 105; mode, 95, 120.*

Median

The median is the "mid-most" value of a data distribution. It is the value above which or below which half of the data points lie.[2,4,5] Alternatively, the median is the 50th percentile value of a distribution.

The median is unaffected by outliers and may be more useful than the mean to describe data when outliers exist[2] or when continuous data are not normally distributed.[4] The median is useful for describing ordinal data[4] because the magnitude of difference between points of a data scale need not be consistent to determine the 50th percentile value.[2] The median is not useful to describe nominal data[2] because of the arbitrary selection of numbers used to generate this scale.

Mode

The mode is the most commonly obtained value or values on a data scale, or the highest point of a peak on a frequency distribution.[2] The mode is most useful when two clusters of data exist (bimodal distribution), such that a group mean is misleading or meaningless.[2] The mode is useful to describe nominal data, defining the most prevalent characteristic of a sample.

Numerical Examples

Three different distributions of data will be examined to determine how the type of data distribution obtained affects the previously defined measures of central tendency.

Figure 1 represents normally distributed data for systolic blood pressure of 30 men aged 31 to 40 years. For normally distributed data, the values of the mean, median, and mode are identical.

Figure 2 presents theoretical systolic blood pressure data of patients with untreated renovascular hypertension. The distribution is negatively skewed. In the absence of normality, the mean, median, and mode are not equal. Also, an outlier, such as a systolic blood pressure value of 150 mm Hg instead of 180 mm Hg, will alter the value of the mean, but not the median or mode.

Figure 3 presents systolic blood pressure for a sample that includes two groups, pregnant women in their second trimester and men. Again, the mean and median are unequal in this non-normally distributed data. Also there exist two peaks of data cluster

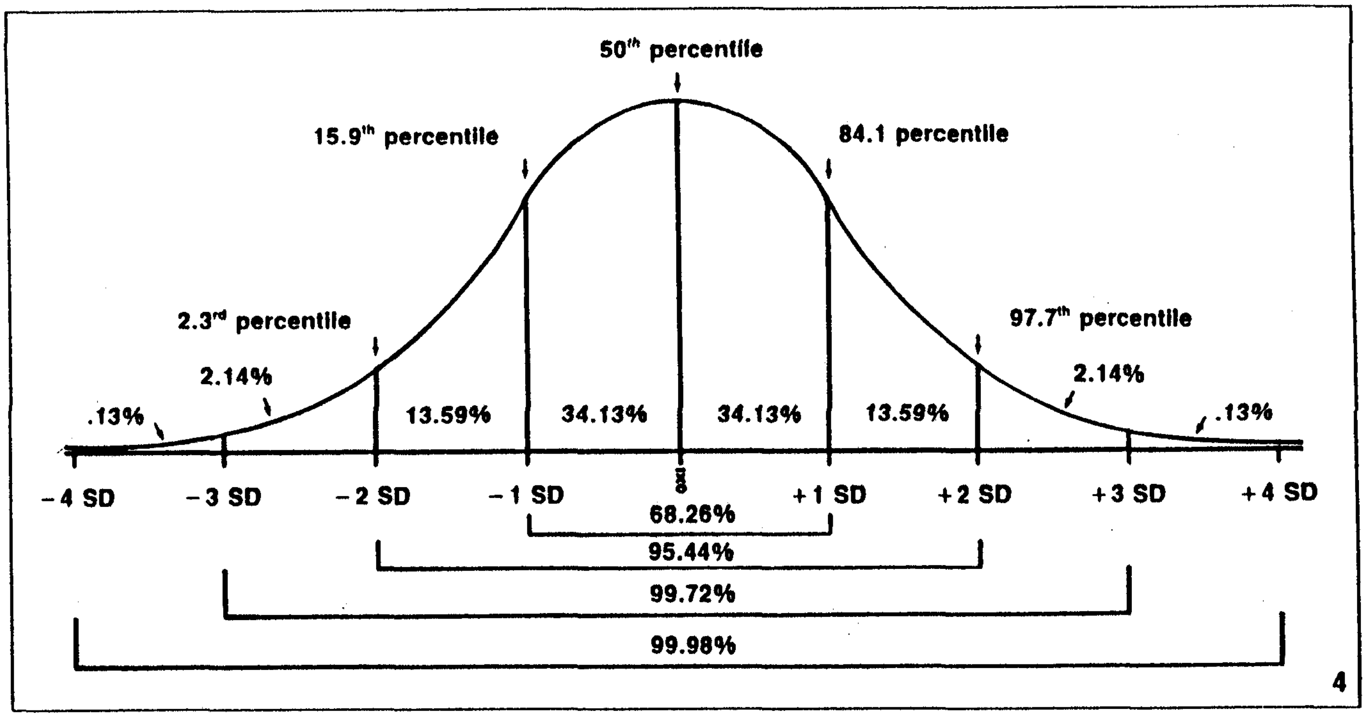


FIGURE 4. *SD and the normal distribution: 68.26% of all scores fall within ± 1 SD from the mean; 95.44% of all scores fall within ± 2 SD from the mean; 99.72% of all scores fall within ± 3 SD from the mean; 99.98% of all scores fall within ± 4 SD from the mean.*[5,6]

TABLE 1. *Applicability of measures of central tendency*

Characteristic	Mean	Median	Mode
Useful with interval, ratio data	Yes	Yes	Yes
Useful with ordinal data	No	Yes	Yes
Useful with nominal data	No	No	Yes
Affected by outliers	Yes	No	No

two modes. To ignore the bimodal aspect of this distribution would be to overlook its unique feature. Also, the presence of an outlier would alter the mean, but not the median or modes.

In Figures 2 and 3 the mean, median, and mode(s) are unequal because data are not normally distributed.[5] Thus, the measure of central tendency most useful to data analysis depends on the type of data, and what aspect of the data is to be conveyed. Fortunately, most physiologic data are normally or near normally distributed so that mean, median, and mode are equal. However, ordinal scale data have no consistent magnitude of difference between units of the data scale, and most ordinal data are not normally distributed.[3] Therefore, the mean is misleading as a measure of central tendency for ordinal scale data.[2,3]

No single measure of central tendency is best for all situations.[5] The applicability of measures of central tendency is summarized (Table 1).

MEASURES OF VARIABILITY

Measures of central tendency do not describe the variability, or spread, of data. Standardized estimates defining data variability are needed to help infer whether two groups studied differ significantly. In other words, measures of variability are used to help infer whether two or more groups studied are drawn from different populations. Several estimates of variability exist.

Range

The range is the interval between the lowest and highest values within a data group.[2] It is the simplest measure of variability to understand and identify. While simple, the range only considers the extreme values of a series of measures, and thus the presence of one outlier can markedly influence the range. The range is purely a descriptive tool and should not be used to infer whether groups differ statistically.

Interquartile Range

The interquartile range is a measure of variability directly related to the median. Recall that the median, a measure of central tendency applicable to ordinal and non-normally distributed data, is the middlemost value of a set of data. The median represents the 50th percentile. The interquartile range is that range described by the interval between the 25th and 75th percentile values.[6]

It has been suggested that the interquartile range be used for describing the variability of data that do not meet parametric analysis standards, such as ordinal scale data.[6] The interquartile range clearly defines where the middle 50% of measures occurs and indicates the spread of the data

TABLE 2. *Estimates of variability of systolic blood pressure data of men aged 31 to 40 years*

Subject	Systolic Blood Pressure	$(\bar{X} - X_i)$	$(\bar{X} - X_i)^2$
1	135	15	225
2	115	5	25
3	110	10	100
4	130	10	100
5	125	5	25
6	125	5	25
7	105	15	225
8	120	0	0
9	120	0	0
10	120	0	0
11	125	5	25
12	110	10	100
13	115	5	25
14	115	5	25
15	135	15	225
16	100	20	400
17	120	0	0
18	125	5	25
19	120	0	0
20	130	10	100
21	140	20	400
22	120	0	0
23	115	5	25
24	110	10	100
25	130	10	100
26	105	15	225
27	120	0	0
28	115	5	25
29	125	5	25
30	120	0	0

Mean = $\Sigma X/n = 3{,}600/30 = 120$
Median = 120
Mode = 120
Variance: $\Sigma (\bar{X} - X_i)^2/(n-1) = 2{,}550/29 = 87.931$
SD = $\sqrt{87.931} = 9.377$
SEM = $\dfrac{9.337}{\sqrt{n}} = 1.712$

without using statistical techniques improperly.

Standard Deviation

The standard deviation (SD) is one of the most commonly encountered estimates of data variability and is integral to performance of inferential statistical techniques.[2] It provides an estimate of the degree of scatter of individual sample data points about the sample mean.

The usefulness of the SD lies in its properties as related to the Gaussian, or normal, distribution. The SD itself can be used to define an extreme score, such as the value that is exceeded by 5% or 95% of all scores from a sample of a population.[2] Figure 4 shows that 68.26% of data points of a normally distributed population fall within plus or minus one SD of the mean, and 95.44% of points fall within plus or minus two SD of the mean.[5]

The SD is calculated as the square root of another term called the variance. Because individual data points will fall both above and below the mean, the effect of direction of differ-

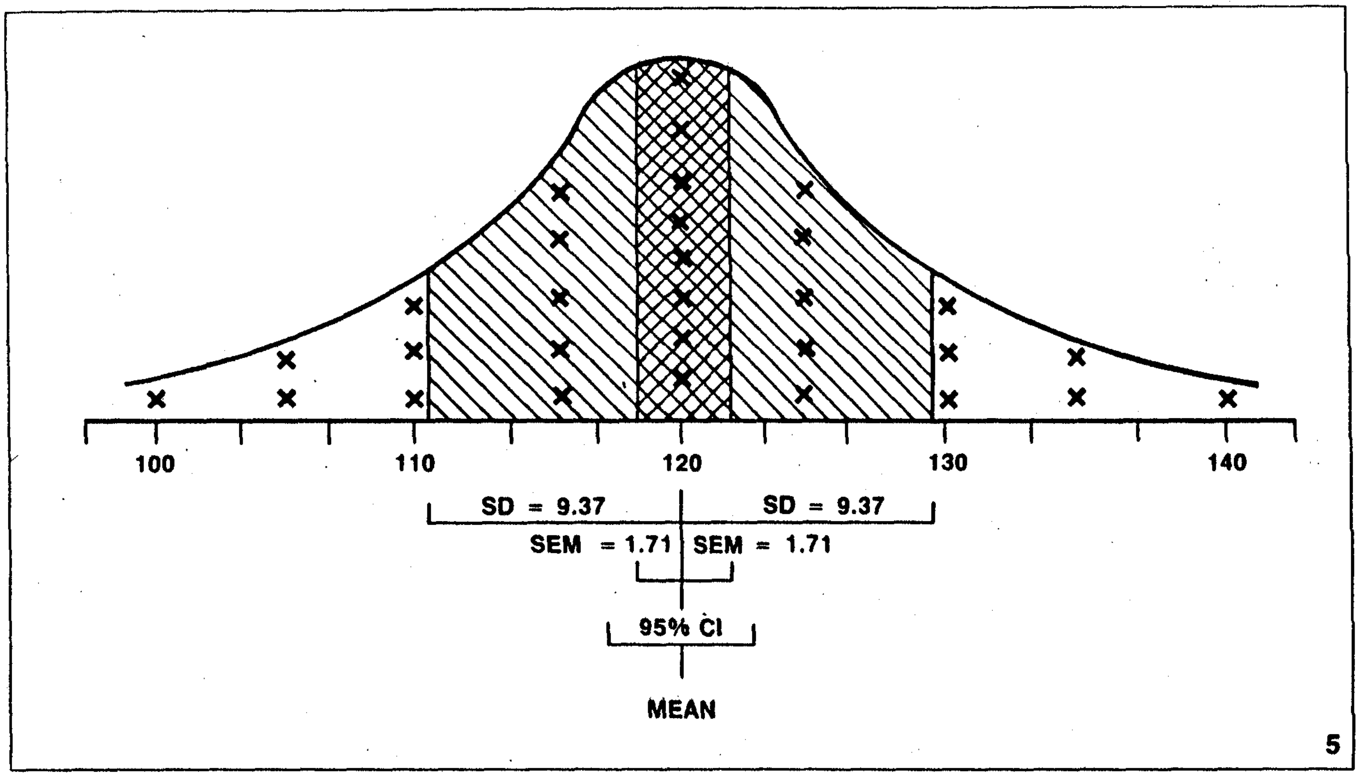


FIGURE 5. *Systolic blood pressure of men aged 31 to 40 years. SD, SE, and 95% CI are noted.*

TABLE 3. *Applicability of measures of variability*

Characteristic	Range	Interquartile Range	SD	SEM
Useful to describe interval or ratio data	Yes	Yes	Yes	Yes
Used to describe ordinal data	Yes	Yes	No	No
Descriptive of sample variability	Yes	Yes	Yes	No
Assists in statistical inference	No	No	Yes	Yes
Used to calculate confidence intervals	No	No	No*	Yes

*SEM = SD/$\sqrt{n}$, thus SD is involved indirectly in calculating a confidence interval.

ence will cause some deviations from the mean to be positive and some to be negative. To overcome this effect, deviations are squared to obtain a positive number. Individual squared deviations from the mean are then averaged to calculate the estimate of variability known as the variance. Numerically

$$\text{Variance} = \Sigma\,(\bar{X} - X_i)^2/(n-1)$$

where $\bar{X}$ equals the mean, X_i equals each individual data point, and n equals the total number of data points.[5-7] The variance represents the deviation from the mean, expressed as the square of the units used. For instance, in Figure 1, the variance of systolic blood pressure is expressed as mm Hg^2. However, these squared units are not meaningful. Therefore, the square root of the variance is then calculated, to bring the variability estimate back to the correct scale. This is the value known as the SD:

$$SD = \sqrt{\text{variance}}$$

The SD is meaningful only when applied to data that are normally or nearly normally distributed.[2,8,9] It is applicable to interval or ratio scale data.[2]

The SD is useful in application to statistical inference techniques. The calculation of SD from the normally distributed systolic blood pressure data of Figure 1 is shown (Table 2).

Standard Error of the Mean

The standard error of the mean (SEM) is a statistic derived from the SD, and is simply calculated as

$$SEM = SD/\sqrt{n}$$

It is obvious from the calculation that the SEM is always smaller than the SD and the greater the n, the smaller the SEM will be.

The SEM is an abstract concept. Imagine repeating an experiment numerous times. With each experiment, a different sample group would be drawn from the study population. Because each repetition of the experiment contains unique sample members, different mean values will be generated with each study. The collection of these mean values, as generated from repetitive sampling and experimentation, will reflect "scatter" about the true but unknown population mean. The SEM is simply a quantification of the variability of these sample mean values. The SEM is properly used to estimate the precision or reliability of a sample, as it relates to the population from which the sample was drawn.[10,11] The SEM does *not* provide an estimate of the scatter of sample data about the sample mean[12] and should not be used as such.

The SEM is useful because it is used in the calculation of "confidence intervals," which contain an estimate of the true mean for an entire population from which the sample was drawn. Confidence intervals can be used for descriptive or inferential purposes.

A calculation of SEM for the normally distributed data presented in Figure 1 is shown (Table 2).

TABLE 4. *Effect of confidence level and sample size on confidence interval width*

Calculation of CIs for Data Presented in Table 2

CI(%)	SD	n	SEM	CI
90	9.377	30	1.712	120 ± 2.82
95	9.377	30	1.712	120 ± 3.36
99	9.377	30	1.712	120 ± 3.83

Effect of Sample Size on CI for Data With A Mean of 120 and a SD of 9.377

CI(%)	SD	n	SEM	CI
95	9.377	30	1.712	120 ± 3.36
95	9.377	100	0.938	120 ± 1.84
95	9.377	,000	0.297	120 ± 0.582

Standard Deviation Versus Standard Error of the Mean

Both SD and SEM are measures of variability. However, the two statistics are different and are frequently confused or misused.[12] The SD defines variability of sample data points about a sample mean. The SD is always greater than the SEM. The SEM is most commonly calculated to help derive confidence intervals.

Various authors have commented about the intellectual sleight of hand of incorrectly using SEM when only SD is appropriate to describe sample data variability.[6,12,13] Bunce et al[13] reviewed 608 articles in six journals in which mean ± SD or SEM were reported. In 50%, SEM values were reported when only the SD would have been appropriate. The authors concluded that "many workers may choose to report the SEM because it is simply smaller than the SD."[13] The inappropriate use of SEM to describe sample data variability may be presented by authors in an attempt to imply that a significant difference exists between groups, when in fact no difference exists. Elenbaas et al[12] were more blunt, concluding that authors who present data as mean ± SEM instead of mean ± SD may be trying to actively impair the reader's ability to accurately identify the variability in the study data.

Whether by error or by design, it is incorrect to underrepresent the variability of sample data as mean ± SEM. We suggest that readers multiply the SEM by $\sqrt{n}$ to obtain the SD when SEM is erroneously used to express sample variability. It is not an error to use SEM in speculating a range, or confidence interval, within which a true *population* mean is likely to fall. The SD and SEM of the data shown in Figure 1 are given (Figure 5). Table 3 summarizes the proper use of estimates of variability.

Confidence Intervals

When statistics derived from the sampling of a population are studied to infer values for population parameters, it would be useful to have confidence that the sample statistical value, such as a mean or SD, would be representative of the true population parameter. One cannot be certain that a sample statistical value is representative of the true population parameter, but one can calculate a range of values likely to be representative of the population parameter.[4,14] That range of values is called a confidence interval (CI). Calculation of a CI is a method of estimating the range of values likely to include the true value of a population parameter. Since one cannot study all members of a population, a representative sample of the population is studied, and from this one uses the mean and SEM to work backward to estimate a CI.

The width of the CI depends on the SEM and the degree of confidence we arbitrarily choose. For instance, a 95% CI, which is the degree of confidence most commonly expressed,[14] is a range of values broad enough that, if the entire population could be studied, 95% of the time the population mean would fall with the CI estimated from the sample of the population.[15] Also; the closer a point lies to the middle of the CI, the more likely it is representative of the population.[16]

Though by convention the 95% CI is most commonly reported, the 95% level is not rigidly required. Wider CIs, such as a 99% or 99.9% CI, are even more likely to include the true population parameter value and are commonly used for critical appraisal of data. They are also advocated for examinations of data during ongoing accumulation of subjects in a clinical trial.[15] Narrower CIs, such as the 90% CI, can be used when study authors find it acceptable that ten times out of 100, the true population parameter may not lie within the CI. However, the width of a CI depends not only on the variability of the data and the level of confidence selected, but also the sample size.

When one broadens a CI by moving a 95% to a 99% CI, accuracy is increased because the calculated CI becomes more likely to include a true population parameter. However, when the level of CI is held constant and sample size is increased, SEM is decreased and thus the CI is narrowed. This narrowing of the CI increases the precision of the CI. The effect of level of confidence selected and sample size on the width of a CI is shown (Table 4).

Calculation of the CI for estimation of true population mean values applies to continuous data from normal or near-normal distributions.[4] Also, a CI can be estimated for such other statistics as medians, regression slopes, relative risk data, response rates, intergroup differences of response rates, X year survival rates, median survival duration, and hazard ratios.[14,16] In addition, CI may be used to visually compare data when two or more sample groups are studied but their members were not randomly selected or assigned between the groups.[15]

Pitfalls in the use of the CI exist. Confidence intervals convey the effects of sampling variation but do not control for such nonsampling errors in study design or execution as improper selection of subjects, poor experimental design, and the like.[4]

SUMMARY

This article has highlighted common proper and improper use of measures of central tendency and measures of variability. The relationship between the type of data scale and the correct use of mean, median, and mode have been presented. The variability estimates of range, interquartile range, SD, and SEM have been introduced, and their proper and improper use has been discussed. Finally, the concept and proper use of CIs have been outlined.

The next installment of this series, in the May issue, will cover hypothesis testing. Included will be types of experimental error, the terms alpha (α) and beta (β), statistical power, and sensitivity, specificity, and predictive value.

REFERENCES

1. Gaddis ML, Gaddis GM: Introduction to biostatistics: Part 1, Basic concepts. *Ann Emerg Med* 1990;19:86-89.

2. Clegg F: Introduction to statistics I: Descriptive statistics. *Br J Hospital Med* 1987;37:356-357.

3. Forrest M, Andersen B: Ordinal scale and statistics in medical research. *Br Med J* 1986;292:537-538.

4. Campbell MJ, Gardner MJ: Calculating confidence intervals for some non-parametric analyses. *Br Med J* 1988;296:1454-1456.

5. Hopkins KD, Glass GV: *Basic Statistics for the Behavioral Sciences.* Englewood Cliffs, New Jersey, Prentice-Hall, Inc, 1978.

6. Glanz SA: *Primer of Biostatistics*, ed 2. New York, McGraw-Hill Book Co, 1987.

7. Sokal RR, Rohlf FJ: *Biometry*, ed 2. New York, WH Freeman and Co, 1981.

8. Maxfield M, Schweitzer J, Gouvier WD: Measures of central tendency, variability, and relative standing in nonnormal distributions: Alternatives to the mean and standard score. *Arch Phys Med Rehabil* 1988;69:406-409.

9. Nierenberg A, Jekel J, Singer B: Is standard deviation always the right choice? (letter). *Am J Psychiatry* 1986;143:1198-1199.

10. Davis RH: How to use standard error in a clinical study. *J Am Podiatric Med Assoc* 1987;77:154-156.

11. Hamer RM: Measures of precision for means: SE or SD? (letter). *Am J Psychiatry* 1986;143:804-805.

12. Elenbaas RM, Elenbaas JK, Cuddy PG, et al: Evaluating the medical literature, Part II: Statistical analysis. *Ann Emerg Med* 1983;12:610-620.

13. Bunce H, Hokanson JA, Weiss GB: Avoiding ambiguity when reporting variability in biomedical data. *Am J Med* 1980;69:8-9.

14. Gardner MJ, Altman DG: Confidence intervals rather than *P* values: Estimation rather than hypothesis testing. *Br Med J* 1986;292:746-750.

15. Bulpitt CJ: Confidence intervals. *Lancet* 1987:494-497.

16. Simon R: Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 1986;105:429-435.

Introduction to Biostatistics: Part 3, Sensitivity, Specificity, Predictive Value, and Hypothesis Testing

Diagnostic tests guide physicians in assessment of clinical disease states, just as statistical tests guide scientists in the testing of scientific hypotheses. Sensitivity and specificity are properties of diagnostic tests and are not predictive of disease in individual patients. Positive and negative predictive values are predictive of disease in patients and are dependent on both the diagnostic test used and the prevalence of disease in the population studied. These concepts are best illustrated by study of a two by two table of possible outcomes of testing, which shows that diagnostic tests may lead to correct or erroneous clinical conclusions. In a similar manner, hypothesis testing may or may not yield correct conclusions. A two by two table of possible outcomes shows that two types of errors in hypothesis testing are possible. One can falsely conclude that a significant difference exists between groups (type I error). The probability of a type I error is α. One can falsely conclude that no difference exists between groups (type II error). The probability of a type II error is β. The consequence and probability of these errors depend on the nature of the research study. Statistical power indicates the ability of a research study to detect a significant difference between populations, when a significant difference truly exists. Power equals 1 − β. Because hypothesis testing yields "yes" or "no" answers, confidence intervals can be calculated to complement the results of hypothesis testing. Finally, just as some abnormal laboratory values can be ignored clinically, some statistical differences may not be relevant clinically. [Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 3, sensitivity, specificity, predictive value, and hypothesis testing. Ann Emerg Med May 1990;19:591-597.]

Gary M Gaddis, MD, PhD*
Monica L Gaddis, PhD†
Kansas City, Missouri

From the Departments of Emergency Health Services* and Surgery,† University of Missouri — Kansas City School of Medicine, Truman Medical Center, Kansas City.

Received for publication September 1, 1989. Accepted for publication January 30, 1990.

Address for reprints: Gary M Gaddis, MD, PhD, Department of Emergency Health Services, University of Missouri — Kansas City School of Medicine, Truman Medical Center, 2301 Holmes, Kansas City, Missouri 64108.

INTRODUCTION

Diagnostic tests guide the physician in assessment of clinical disease entities. In a similar manner, statistical inference theory guides the scientist in the testing of scientific hypotheses. Before discussing inferential techniques (parts 4 and 5 of this series), it is necessary to understand the basis of hypothesis testing, to gain an appreciation of the type of questions inferential statistics help answer. Clinical diagnostic testing and hypothesis testing have many parallels, but most clinicians are more familiar with diagnostic than hypothesis testing. Therefore, this article will focus on the components of diagnostic testing theory, including sensitivity, specificity, and predictive value. This will be followed by analogies to facilitate understanding of hypothesis testing.

EVALUATION OF DIAGNOSTIC TESTS

Sensitivity and Specificity

Physicians make medical diagnoses with the aid of the patient history, physical examination, and diagnostic testing. Numerous new diagnostic tests are presented each year in the medical literature, and each must be evaluated before it is introduced into the clinical setting. Most new diagnostic tests are evaluated in relation to another older, previously accepted, often more invasive, and historically reliable test (the "gold standard" test). Common examples of gold standards include the use of ECG changes plus cardiac enzyme levels to diagnose acute myocardial infarction, or pulmonary angiography to diagnose pulmonary embolism. For the purposes of our discussion, it will be assumed that results obtained by the gold standard test are always correct.

Hypothetically, imagine that a new magnetic resonance imaging (MRI) venogram has been proposed as a noninvasive means of evaluating patients suspected by clinical criteria of having a deep venous thrombosis. The MRI venogram, the proposed new diagnostic test, will be evaluated against the traditional and widely used gold standard, the intravenous contrast venogram. Table 1 shows that there are four possible outcomes of diagnostic testing. Patients can be diagnosed as having deep venous thrombosis or not having deep venous thrombosis by both the gold standard test and by the new MRI diagnostic test, if patients undergo both tests.

In Table 1, 250 patients clinically suspected of having deep venous thrombosis undergo both tests. Of the 250 patients clinically suspected to have deep venous thrombosis, 150 actually do have deep venous thrombosis, with 130 shown to have deep venous thrombosis by both the gold standard test and by the new MRI test. This group of 130 is termed the true positive (TP) group by the new diagnostic test because they are shown to have disease by the new test and are also proven to have disease by the gold standard test. However, 20 of the 150 patients who are proven by the gold standard test to have deep venous thrombosis had a negative MRI diagnostic test. These 20 are termed the false negative (FN) group because they were classified incorrectly as disease free by the new MRI test.

Similarly, 100 of the patients were judged disease free by the contrast venogram, but of these, only 87 had a negative MRI test. This group of 87 constitutes the true negative (TN) group. The remaining 13 were incorrectly classified by the new MRI test as having a deep venous thrombosis, when in fact they did not have the disease. This constitutes the false positive (FP) group.

The two by two outcome table in Table 1 can now be used to help us evaluate how well the new MRI test does in detecting deep venous thrombosis. We want to know the answers to two questions: Is the test sensitive enough to detect the presence of a deep venous thrombosis in a diseased patient? Is the test specific enough to indicate the absence of deep venous thrombosis disease only in patients who in fact are not afflicted by it?

TABLE 1. *Gold standard versus diagnostic test*

		Gold Standard Test (Contrast Venogram) Disease Evident	No Disease Evident	Total
Diagnostic Test	Disease Evident	TP (130)	FP (13)	143
(MRI Venogram)	No Disease Evident	FN (20)	TN (87)	107
		150	100	250

TABLE 2. *Gold standard versus diagnostic test*

		Gold Standard Test (Contrast Venogram) Disease Evident	No Disease Evident	Total
Diagnostic Test	Disease Evident	TP (35)	FP (21)	56
(MRI Venogram)	No Disease Evident	FN (5)	TN (139)	144
		40	160	200

TABLE 3. *Possible outcomes of hypothesis testing*

	Reality H_0 False, H_1 True	H_0 True, H_1 False
Decision From Statistical Test		
Reject H_0, Accept H_1	Correct, No Error (A)	Incorrect, Type I Error (B)
Accept H_0, Reject H_1	Incorrect, Type II Error (C)	Correct, No Error (D)

Sensitivity, which can be thought of as "positivity (of the test) in disease," is derived by working down the first column of Table 1:

$$\text{Sensitivity (\%)} = 100 \times TP/(TP + FN)$$

In this example, sensitivity equals 100 x 130/(130 + 20), or 86.7%.

Specificity, which can be thought of as "negativity (of the test) in health," is also derived by working vertically, in the second column of Table 1:

$$\text{Specificity (\%)} = 100 \times TN/(TN + FP)$$

Here, specificity equals 100 x 87/(87 + 13), or 87.0%.

The ideal diagnostic test would be 100% sensitive and 100% specific, and thus would have no FP or FN

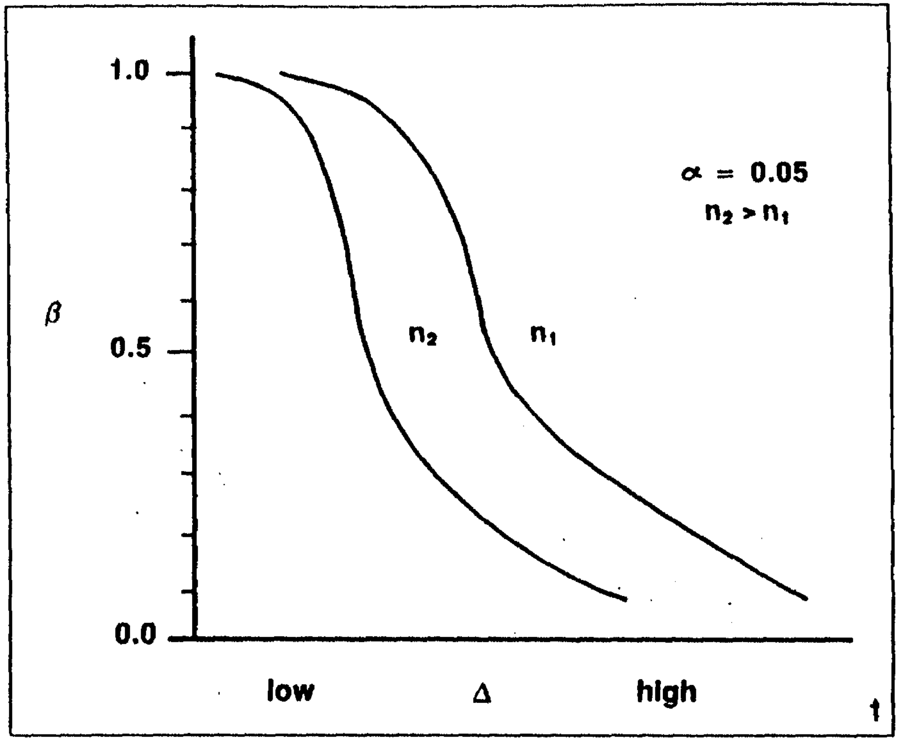


FIGURE 1. *Operating characteristic curve. β is dependent on α, n, and Δ. In this example, α is fixed at .05. All else held constant, increasing Δ or increasing n decreases β.*

TABLE 4. *Prior probability and chance of error*

	Prior Probability	
	Low	High
Chance of Error		
Type I	High	Low
Type II	Low	High

outcomes. Because virtually all diagnostic tests have some FP and FN outcomes, they do not have 100% sensitivity and specificity.

Unfortunately, many clinicians believe that sensitivity and specificity can be used to predict whether an individual patient is diseased or disease free. This is an error. Sensitivity and specificity are merely properties of a test. Sensitivity and specificity should not be used to make predictive statements about an individual patient.

Predictive Value

Predictive values can be used to help predict the likelihood of disease in an individual. A positive predictive value (PPV) is useful to indicate the proportion of individuals who actually have the disease when the diagnostic test indicates the presence of that disease. A negative predictive value (NPV) is useful to determine the proportion of individuals who are truly free of the disease tested for when the diagnostic test indicates the absence of that disease.

Predictive values are derived by working horizontally on the two by two outcome table in Table 1:

$$\text{PPV } (\%) = 100 \times \text{TP}/(\text{TP} + \text{FP})$$
$$\text{NPV } (\%) = 100 \times \text{TN}/(\text{TN} + \text{FN})$$

From the example in Table 1, PPV = 100 x 130/(130 + 13), or 90.9%, and NPV = 100 x 87/(87 + 20), or 81.3%.

PPV and NPV are affected by the prevalence of disease in the population. Prevalence is defined as the proportion of the population afflicted by the disease in question. In the example in Table 1, the prevalence of deep venous thrombosis when it was clinically suspected was 60% because the total number of patients studied was 250, and the number of patients who actually had a contrast venogram (the gold standard test) indicative of deep venous thrombosis was 150.

Next, the effects of decreased prevalence of deep venous thrombosis on the predictive value of the MRI venogram test will be examined. Imagine a sample of 200 patients, only 20% of whom have a deep venous thrombosis (prevalence, 20%). This group is depicted (Table 2). Because 20% of the patients have a deep venous thrombosis, the sum of TP + FN in column 1 must be 0.2 x 200, or 40. Of these, about 35 will constitute the TP group because the sensitivity of the test has already been shown to be 86.7% (0.867 × 40 = 34.7). The remaining five can be expected to be in the FN group because sensitivity is a property of the test independent of disease prevalence. Because the prevalence of deep venous thrombosis is only 20%, the remaining 0.8 x 200, or 160, will not have a deep venous thrombosis, so the sum of TN + FP results in column 2 will be 160. Of this set of 160, 87%, or about 139, will be in the TN group, and the remaining 21 will be in the FP group because specificity is also a property of the test, independent of disease prevalence.

The change of prevalence markedly influences the PPV and NPV values obtained (Table 2). With a 20% prevalence, the PPV falls to 100 x 35/(35 + 21), or 62.5%, while the NPV increases to 100 x 139/(139 + 5), or 96.5%. Note that as disease prevalence falls, the PPV of any test will fall and the NPV of any test will increase.

From this, it is easy to see why many new diagnostic tests that seem from initial reports to be useful may not represent a diagnostic improvement when in common use. Many diagnostic tests are validated in settings on populations with a high prevalence of the disease for which testing is done. However, when the new test is used in different clinical settings with a lower prevalence of

FIGURE 2. *Clinical testing.*

Sensitivity	The ability of a test to reliably detect the presence of disease (positivity in disease). Sensitivity (%) = 100 × TP/(TP + FN)
Specificity	The ability of a test to reliably detect the absence of disease (negativity in health). Specificity (%) = 100 × TN/(TN + FP)
Prevalence	The proportion of the population with disease. Prevalence (%) = 100 × (TP + FN)/(n)
Positive Predictive Value	The proportion of individuals with disease when the presence of disease is indicated by the diagnostic test. PPV = 100 × TP/(TP + FP)
Negative Predictive Value	The proportion of individuals free of disease when the absence of disease is indicated by the diagnostic test. NPV = 100 × TN/(TN + FN)

TN, true negative; FN, false negative; TP, true positive; FP, false positive.

2

that disease, the test does not perform up to reported expectations. A clinical example of the interrelationship between prevalence of disease and predictive value is the use of amylase levels to screen for pancreatitis. An elevated amylase level is more likely indicative of pancreatitis in persons previously afflicted with pancreatitis than it is predictive of pancreatitis among all patients with abdominal pain or other possible causes of an elevated serum amylase level.

In summary, sensitivity and specificity are properties that indicate the degree of reliability of a diagnostic test. Sensitivity and specificity do not indicate predictive value. Predictive values can be applied to an individual patient's test result and are affected by the prevalence of the disease in the population to which the test is applied. The PPV will fall and the NPV will rise as the prevalence of disease decreases.

HYPOTHESIS TESTING

Formulation of the Hypothesis

Statistical inference involves the testing of hypotheses. A hypothesis is a numerical statement about an unknown parameter.[1] Just as a two by two table can be constructed for the four possible outcomes of a clinical diagnostic test, a two by two table can be constructed for the four possible outcomes of hypothesis testing. '

Before constructing this table, it is necessary to understand what a hypothesis states. The first step in hypothesis testing is a statement of a hypothesis in positive terms. This defines the "research" or "alternative" hypothesis, H_1.[2] For example, one could hypothesize that experienced emergency physicians (those with more than five years of full-time postgraduate emergency department experience) can examine, diagnose, and treat more patients per hour than inexperienced emergency physicians (less than five years of full-time ED experience).

The next step is to state the "null" or "statistical" hypothesis, H_0, which follows logically from H_1.[1,2] The hypothesis tested statistically is H_0. In this example, H_0 would state "Experienced emergency physicians and inexperienced emergency physicians do not differ significantly in the number of patients they can examine, diagnose, and treat per hour."

We "reject" or "fail to reject" ("accept") H_0 based on our inferential statistical testing.[1-3] H_0 hypothesizes a difference of zero between population samples tested, while H_1 hypothesizes a nonzero difference between population samples tested. There exist an infinite number of possible nonzero differences between populations. Therefore, the reason that H_0 rather than H_1 is tested is that mathematically, H_0 theorizes a single magnitude of difference between populations studied, and it is possible to statistically assess this single hypothesis. In contrast, H_1 is actually an infinite number of hypotheses because there exist an infinite number of possible magnitudes of difference between populations.[4] It would be impossible to calculate the required statistics for each of the infinite number of possible magnitudes of difference between population samples H_1 hypothesizes.

If H_0 is "accepted" as tenable, then H_1 must be "rejected," and vice versa, because the two hypotheses are mutually exclusive. When H_0 is tested, the probability that numerical differences between population samples are not due strictly to chance is assessed.[2] H_0 does recognize that nonzero differences between groups are possible, even if two samples of the same population are tested, simply due to random scatter of the data.[2] If H_0 is "accepted" as tenable, this signifies the likelihood that no significant difference exists between the populations studied and that any numerical differences between groups are due to chance alone. If H_0 is rejected, this signifies that a significant difference does exist between the populations studied and that the numerical differences between the groups are not due to chance alone.

Errors in Hypothesis Testing

Hypothesis testing may lead to erroneous inferential statistical conclusions, just as diagnostic testing may lead to erroneous diagnostic conclusions. Just as a two by two table of possible outcomes of diagnostic tests can be constructed, so can a two by two table of possible outcomes of inferential statistical tests be constructed (Table 3). Two types of incorrect conclusions are possible. Box B of Table 3 indicates cases in which the statistical test falsely indicates that a significant difference exists between groups, when in fact no true difference exists. It is analogous to a false-positive diagnostic test result. In other words, box B shows cases where H_0 is rejected, when it is in fact true. This rejection of H_0 when H_0 is true is arbitrarily called a type I error.[1-3]

Box C of Table 3 indicates cases in which the statistical test falsely indicates the lack of a significant difference between groups, when in fact a true difference exists (H_1 is true). This is analogous to a false-negative diagnostic test result. In other words, box C shows cases in which H_0 is accepted when it is in fact false. The acceptance of H_0 when H_0 is false is arbitrarily called a type II error.[1-3]

FIGURE 3. *Hypothesis testing.*

Research (Alternative) Hypothesis (H_1)	An hypothesis that states a difference exists between two (or more) populations studied. H_1 is a positive statement that a difference exists between groups.
Null (Statistical) Hypothesis (H_0)	An hypothesis of no difference between two or more populations studied. H_0 is a negative statement, that no difference exists between groups.
Type I Error	To reject the null hypothesis (H_0), when in fact H_0 is true. To falsely conclude that a significant difference exists between populations.
Type II Error	To accept the null hypothesis (H_0), when in fact H_0 is false. To falsely conclude that no significant difference exists between populations.
Alpha (α)	The probability of making a type I error.
$P < .05$	Statistical calculations from the experimental data indicate that the probability of making a type I error is less than 5%.
Beta (β)	The probability of making a type II error.
Power	The ability of an experiment to find a significant difference exists between populations, when in fact a significant difference truly exists. Power = $1 - \beta$
Delta (Δ)	The degree of difference between populations tested.
Operating Characteristic Curve	A function that relates the dependent variable β that results from independent values of α, Δ, and n.
Prior Probability	The likelihood that an hypothesized difference between populations is in fact correct.

3

Box A and box D of Table 3 denote correct conclusions, analogous to true-positive and true-negative diagnostic test results. Thus, Table 3 shows that there exist two correct and two incorrect conclusions possible whenever H_0 is tested.

Next, the probability of making incorrect conclusions must be assessed. The probability of making a type I error is defined as alpha (α).[1,2,4] α is derived from the raw data, statistical calculations, and statistical tables appropriate for the inferential statistical test used. By convention, statistical significance is generally accepted if the probability α of making a type I error is less than 0.05, which is commonly denoted on figures and tables as $P < .05$.[3,4]

Though conventional, selection of an alpha level of .05 as the crucial level of significance is arbitrary. Accepting significance at $\alpha = .05$ means that it is recognized that one time out of 20, a type I error will be committed, a consequence that the investigator is willing to accept. If the consequences of making a type I error are judged to be sufficiently severe, it may be appropriate to select more stringent levels of α, such as .01, as the cutoff for statistical significance. When a caption or text indicates that for some statistical comparison, $P = .XY$, the probability of a type I error, based on the calculations performed for that inferential statistical test, is 0.XY, and the reader is left to judge whether this level of α is indicative of a true difference between populations tested. Another advantage of the reporting of *P* values is that the arbitrary designation of significance at .05, and the improper and arbitrary designation of a trend if $.10 > P > .05$, can be avoided.

The probability of making a type II error is defined as beta (β).[1,2,4] β is more difficult to derive than α, and unlike α, actually is not one single probability value. β is often ignored by researchers.[5] However, it is important. If some treatment yields a 10% increase in survival or a 10% decrease in some complication, it would likely be readily incorporated into medical practice. Unfortunately, numerous clinical trials have suffered from errors of experimental design that cause β to be unacceptably high, such that type II errors are easily made, and treatments that are significantly better than older methods are rejected because of statistical artifact resulting from poor experimental design.[5] By convention, β should be less than .20, and ideally less than .10, to minimize the chance of making a type II error.[6]

α and β are interrelated. All else held constant (such as the populations studied, the number of subjects, and the method of testing), as α is arbitrarily decreased, β is increased. As α is increased, β is decreased.[1,2]

Statistical power is defined as $(1-\beta)$.[1,2,4] Because β indicates the probability of making a type II error, power indicates mathematically the probability of not making a type II error. Power is analogous to sensitivity in hypothesis testing. Sensitivity indicates the probability that the diagnostic test can detect disease when it is present. Power indicates the probability that the statistical test can detect significant differences between populations, when in fact such differences truly exist.

Power depends on several variables:[1,2,4,7]

α: As α increases, β decreases, and power increases.

n (sample size): As n increases, power increases.

The magnitude of the difference actually present between the populations tested, delta (Δ): Just as it is easier to find a pitchfork than a needle in a haystack, so it is easier to find a large difference than it is to find a small difference between populations tested.

One-tailed versus two-tailed tests: One-tailed tests are more powerful than two-tailed tests, because a statistical test result must not vary as much from the mean to achieve significance at any level of α chosen. (If α is .05, for a two-tailed test, a result must fall in either the top or bottom 2½% of results to achieve significance, but for a one-tailed test, the result must merely fall in either the top or bottom 5% of a distribution.) In the original hypothesis example about how quickly emergency physicians can treat patients, the appropriate test would be one-tailed, because H_1 specifies the direction of the difference between groups hypothesized.

Parametric versus nonparametric statistical testing: Parametric tests are generally more powerful. (This will be further discussed in Part 4 of this series.)

Use of proper experimental design and statistics: Errors in these areas decrease power.

Because so many variables can affect β, β is not one single value. This follows from the fact that α is the probability of erroneously concluding that H_0 is false, and H_0 specifies a single magnitude of difference between populations. However, as has been explained, β is the probability of erroneously concluding that H_1 is false, and H_1 hypothesizes an infinite number of possible magnitudes of difference between populations tested. β is expressed as a function of $\triangle$, n, and α by a function called the operating characteristic curve of the test[5] (Figure 1).

The most common use of β is in the calculation of the approximate number of subjects that must be studied to keep α and β acceptably small. This calculation uses estimates of population standard deviations and estimates of $\triangle$, acceptable values of α and β, and numbers from statistical tables, to derive a value of n of sufficient size. The determination of adequate sample size for an experiment is readily referenced.[8-10]

P Values Versus Confidence Intervals

Hypothesis testing yields yes or no answers about statistical significance, answers that can be fraught with errors, and answers that may represent oversimplifications. *P* values imply little about the magnitude of difference present between populations. Therefore, some feel that the use of confidence intervals (CIs) is complementary or even preferable to the use of *P* values in reporting clinical data.[11] (Confidence intervals were discussed in part 2 of this series.[12]) It is correct to report both CI and *P* values for scientific data, and the two are often complementary.[1,11]

Clinical Versus Statistical Significance

Statistically significant numerical differences between study groups may not be clinically significant or relevant. An analogy to clinical testing is again useful. It is common experience to ignore or place little emphasis on a single diagnostic test result that lies outside the expected range for that test when large numbers of tests are done. An example is the interpretation of an isolated elevated amylase level in a patient having otherwise normal routine laboratory data after a normal screening physical examination at his family physician's office. Many experienced clinicians can intuitively sense when to place little emphasis on isolated laboratory test results outside the normal range when an abnormal result is not expected. Alternatively stated, when there is very little prior probability of disease, an isolated abnormal laboratory value is generally not cause for great concern, and the clinician avoids a clinical error analogous to a type I error by avoiding concluding that disease is present in a disease-free patient.

Similarly, if enough statistical comparisons are made, eventually type I and type II statistical errors are inevitable. The problem comes in discerning which statistically significant differences are meaningful and which are meaningless. Just as prevalence affects the predictive value of a positive diagnostic test, so the prior probability of a difference affects the predictive value of a statistical test. Prior probability is an expression of how likely an hypothesis will be true when assessed *before* doing statistical calculations. Prior probability is derived from previously available knowledge that led to the formulation of the hypothesis being tested.

When a hypothesis has a low prior probability of being true, yet achieves statistical significance, such as a link between coffee consumption and pancreatic cancer,[13] a significant result must be interpreted cautiously. Furthermore, if a type I error is being made, repetitive study will probably not replicate a significant difference, as subsequently occurred in the case of the alleged link between coffee consumption and pancreatic cancer.[14] However, in cases of high prior probability, a significant statistical difference is usually correct, just as in cases of high disease prevalence, a positive clinical test result is more likely to be correct.

Table 4 summarizes the interrelationship between prior probability and the chance of making a type I or type II error. This relationship is further explained by Bayes theorem, which the reader is invited to explore.

SUMMARY

An understanding of the interpretation of diagnostic tests facilitates an understanding of hypothesis testing. A diagnostic test result may be a true-positive, true-negative, false-positive, or false-negative result. For diagnostic tests, sensitivity and specificity are properties of the diagnostic test and do not indicate predictive value. Prevalence of disease is a determinant of the predictive value of both positive and negative test results.

Similarly, hypothesis testing can yield erroneous results. A false-positive result, which accepts the presence of a significant difference between populations when in fact no significant difference exists (type I error), occurs with a probability of α. A false-negative result, rejecting the presence of a significant difference between populations, when in fact they actually do differ (type II error), occurs with a probability of β.

Power is 1-β, and is analogous to the sensitivity of a diagnostic test in that both sensitivity and power address whether a test can detect what it is designed to detect. As sensitivity and specificity are not predictive, so also power is not predictive. As prevalence of disease affects the predictive value of a positive test result, so the prior probability of a difference being present affects the predictive value of a significant statistical test result. Figures 2 and 3 summarize these points.

REFERENCES

1. Hopkins KD, Glass GV: *Basic Statistics for the Behavioral Sciences.* Englewood Cliffs, New Jersey, Prentice-Hall, Inc, 1978.

2. Keppel G: *Design and Analysis. A Researcher's Handbook.* Englewood Cliffs, New Jersey, Prentice-Hall, Inc, 1978.

3. Elenbaas RM, Elenbaas JK, Cuddy PG: Evaluating the medical literature Part II: Statistical analysis. *Ann Emerg Med* 1983;12:610-620.

4. Sokal RR, Rohlf FJ. *Biometry* (ed 2). New York, WH Freeman and Co, 1981.

5. Freiman JA, Chalmers TC, Smith H, et al: The importance of beta, the type II error, and sample size in the design and interpretation of the randomized clinical trial. *N Engl J Med* 1978;299:690-694.

6. Reed JF, Slaichert W: Statistical proof in inconclusive "negative" trials. *Arch Intern Med* 1981;141:1307-1310.

7. Cohen J: Differences between proportions, in: *Statistics in Medicine*. Boston, Little, Brown, & Co, 1974.

8. Arkin CF, Wachtel MS: How many patients are necessary to assess test performance? *JAMA* 1990;263:275-278.

9. Fleiss JL: *Statistical Methods for Rates and Proportions* (ed 2). New York, John Wiley & Sons, 1981.

10. Young MJ, Bresnitz EA, Strom BL: Sample size nomograms for interpreting negative clinical studies. *Ann Intern Med* 1983;99:248-251.

11. Gardner MJ, Altman DG: Confidence intervals rather than P values: Estimation rather than hypothesis testing. *Br Med J* 1986;292:746-750.

12. Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 2, descriptive statistics. *Ann Emerg Med* 1990;19:309-315.

13. MacMahon B, Yen S, Trichopoulos D, et al: Coffee and cancer of the pancreas. *N Engl J Med* 1981;304:630-633.

14. Gorham ED, Garland CF, Garland FL, et al: Coffee and pancreatic cancer in a rural California county. *West J Med* 1988;148:48-51.

SPECIAL CONTRIBUTION
biostatistics

Introduction to Biostatistics: Part 4, Statistical Inference Techniques in Hypothesis Testing

Statistical methods used to test the null hypothesis are termed tests of significance. Selection of an appropriate test of significance is dependent on the type of data to be analyzed and the number of groups to be compared. Parametric tests of significance are based on the parameters, mean, standard deviation, and variance, and thus are used appropriately when interval or ratio data are analyzed. The t*-test and analysis of variance (ANOVA) are examples of parametric tests of significance. Assumptions regarding the data to be analyzed when using the* t*-test or ANOVA include normality of the populations from which the sample data are drawn, homogeneity of the variances of the populations from which the sample data are drawn, and independence of the data points within a sample group. The* t*-test is the appropriate test of significance to use if there are only two groups to compare. If there are three or more groups to compare, ANOVA is the appropriate test. ANOVA holds the preset α level constant. While ANOVA will imply a significant difference between the groups compared, a multiple comparison test will define which of the three or more groups differ significantly. [Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 4, statistical inference techniques in hypothesis testing. Ann Emerg Med July 1990;19:820-825.]*

Gary M Gaddis, MD, PhD*
Monica L Gaddis, PhD†
Kansas City, Missouri

From the Departments of Emergency Health Services* and Surgery,† Truman Medical Center, University of Missouri-Kansas City School of Medicine.

Received for publication September 1, 1989. Accepted for publication March 30, 1990.

Address for reprints: Monica L Gaddis, PhD, Department of Surgery, Truman Medical Center, 2301 Holmes, Kansas City, Missouri 64108.

INTRODUCTION

The research process follows an organized, stepwise pattern. A problem is identified, the research hypothesis is generated, methods of data collection are devised, and the statistical analysis of the data to be collected is designed. Calculation of measures of central tendency and variability are easily completed, but alone these numbers have only descriptive value. Making a decision to reject or accept the null hypothesis (H_0) requires much more extensive statistical analysis of the data.

Statistical methods used to test the null or statistical hypothesis (H_0) are termed tests of significance.[1] Recall from Part 3 of this series [May 1990;19:591-597] that hypothesis testing involves accepting or rejecting H_0.[2] Selection of an appropriate test of significance is dependent on several factors, including the number of groups to be compared and the type of data to be analyzed. This fourth in the series of six articles will address the concepts of parametric statistical inference techniques in hypothesis testing.

PARAMETRIC VERSUS NONPARAMETRIC METHODS

The mean and the standard deviation (SD) of a population describe a normally distributed population.[3] (Because the SD is computed as the square root of the variance, it can also be said that the variance also describes a normal distribution.) Not only are the mean, median, and mode equal in a normal distribution of data, but known percentages of data fall within set SDs from the mean with a normally distributed set of data. The mean, SD, and variance of a population are termed parameters of that population. Parametric statistical methods are based on these parameters.[1] Thus, given the relationship between these parameters and normality, the underlying assumption of parametric statistical methods is that the data being analyzed are normally distributed. If the data are not normally distributed and cannot be defined as interval or ratio data, other statistical

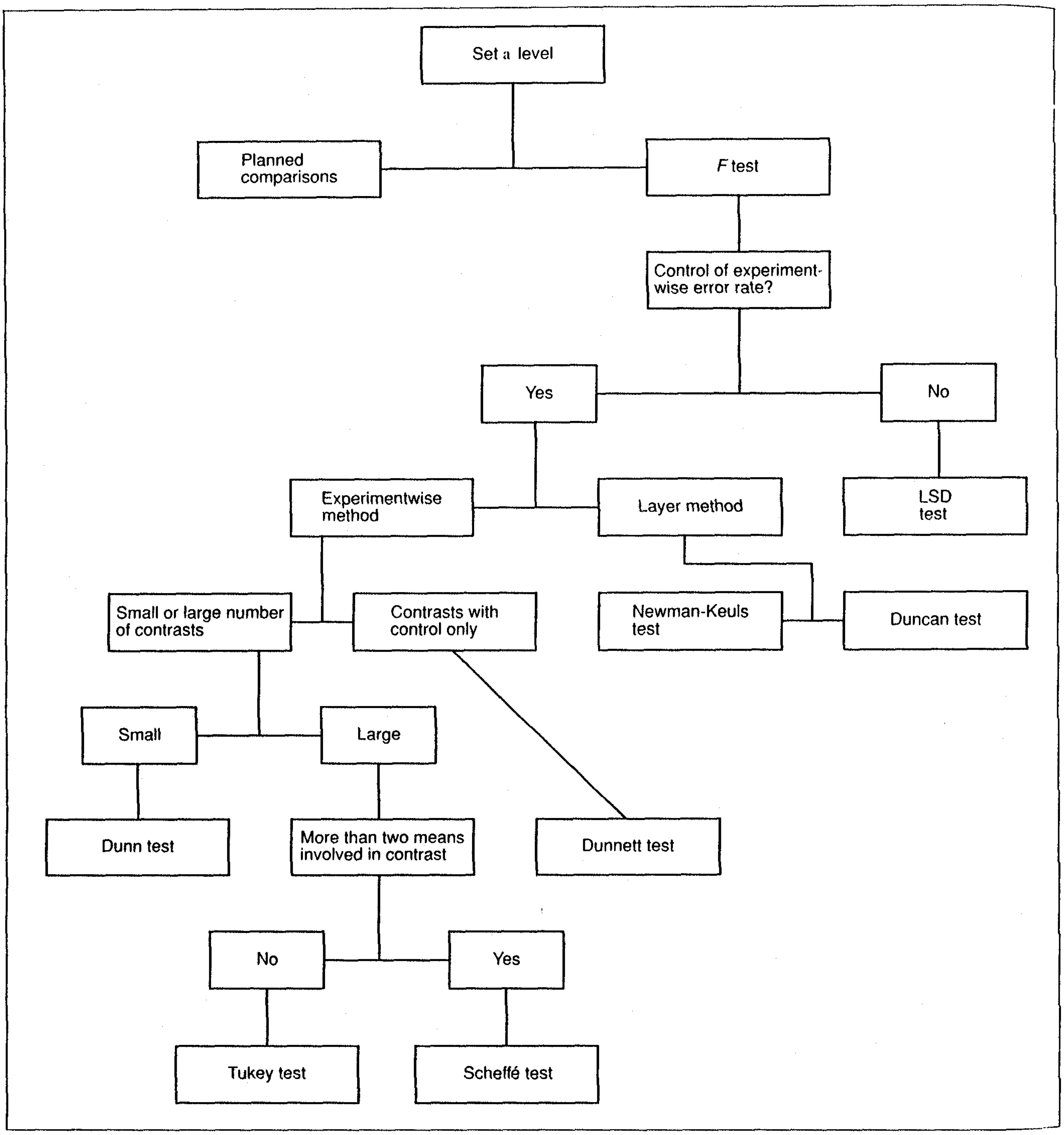


FIGURE. *Flow chart for multiple comparison decisions (adapted from Hopkins and Chadbourn [1967] and Keppel [1973]).*

methods appropriately termed nonparametric statistical methods are used.

In addition to differences in type of data analyzed and the assessment of normality of the data, there are other characteristics possessed by these two classifications of statistical tests that illustrate their inherent differences. First, parametric tests prove to be more powerful than nonparametric tests. That is, if a difference between groups truly exists, all else being the same, that difference would more likely be found using the parametric test. Furthermore, more information about the data is generated from parametric tests.[1] However important these differences are, the nonparametric statistical test should not be discounted. Because not all data are normally distributed and not all are of an interval or ratio scale, nonparametric methods that are sound in their mathematical theory often offer the only legitimate means of data analysis available.

PARAMETRIC STATISTICAL INFERENCE TESTS

t-Test

Student's *t*-test (*t*-test) is the parametric statistical method with which researchers are most often familiar. It is certainly the most common statistical method reported in the medical literature.[1] The *t*-test is used to accept or reject H_0. It is simplistic in that a comparison between two groups can be made and a decision rendered without further analysis. Yet the *t*-test is powerful; it is a parametric method that mathematically and theoretically is based on the means, SDs, and variances of the data.

The *t*-test also requires that several assumptions regarding the data be made prior to use. If the data do not meet the assumptions, then the *t*-test is not the appropriate method to use. Assumptions of the *t*-test include the following: 1) The populations from which the samples were drawn should approach a normal distribution; 2) the variances of the populations from which sample 1 and sample 2 were drawn should be equal or nearly equal; and 3) the observations *within* a population or sample group should be independent, ie, "not paired, matched, correlated, or interdependent in any way."[4]

While these assumptions are important, the *t*-test is robust enough to be an appropriate test if an assumption is not met in the strictest sense (excepting the assumption of independence, which must be met at all times).[4,5] However, this is not to say that it is appropriate to use the *t*-test for nominal or ordinal data or data that do not come from a normally or near-normally distributed population.

While the *t*-test is used to compare two sample groups, the experimental design of the study must be considered because not all *t*-tests are the same. Consideration of the following is important: 1) Are the observations *between* groups independent (as is the case for a control vs experimental group design), so that a *nonpaired* *t*-test is appropriate? 2) Are the observations *between* groups dependent (as is the case for a pretest/post-test design), so that a *paired* *t*-test is appropriate? 3) Are the groups equal or unequal in size? 4) Is the comparison between a population mean and sample mean or between two sample means? 5) Is the direction of the difference between the two groups known or unknown? If a direction of difference is postulated, the *t*-test is termed a one-tailed test. If no direction of difference is postulated, the *t*-test is termed two-tailed.

A very common experimental design in the medical literature is a situation in which there are two different independent groups, a control group and an experimental group.

For example, suppose a new drug is being tested to see if it will decrease arterial pressure in persons with hypertension. Two sample groups would be selected by random assignment. Group 1 will receive a placebo while group 2 will receive the drug in question. The alpha (α) level is preset. (Because the drug in question is hypothesized to lower arterial pressure, a direction of change is postulated, and this data should be tested by a *one-tailed* *t*-test.) The data are collected, descriptive statistics are calculated, and the *t* value is computed. The *t*-test calculation is easily referenced.[4-6]

Once a *t* value is obtained, the researcher should consult a table of critical values for *t* with the appropriate α level and degrees of freedom. If the calculated *t* value is greater than the critical *t*, H_0 is rejected and it is concluded that the medication in question does lower diastolic arterial pressure in hypertensives. If the calculated *t* value is less than the critical *t*, H_0 is accepted as tenable.

Another experimental design common to the medical literature is the pretest/post-test design. This results in dependent or related data between groups (repeated measure) and is analyzed using the *paired* *t*-test.

For example, a new thrombolytic agent is developed that is postulated to halt the progression of a myocardial infarction. Patients entering the emergency department with an evolving myocardial infarction undergo Doppler echocardiography to assess ejection fraction. Following this procedure, the experimental thrombolytic agent is administered. Two days later, ejection fraction is assessed again. Pre- and post-thrombolytic administration data are compared using a paired *t*-test so that patients serve as their own controls. The lack of a significant difference between pre- and post-treatment ejection fraction estimates is expected if the drug is efficacious.

The *t*-test is the method of choice when making a single comparison between two groups whose data meet the assumptions required of parametric analysis methods. However, what is done if the experimental design consists of three or more groups to be compared? The researcher may incorrectly compare these groups using several *t*-tests. For example, if an experiment consisted of one control group (C), and three experimental groups (E1, E2, E3), the comparisons made using *t*-tests would be C versus E1, C versus E2, C versus E3, E1 versus E2, E1 versus E3, and E2 versus E3. While this seems logical and certainly easy, it is improper and can lead to serious errors in drawing conclusions from the data.[1,4-6]

When several groups from an experiment are compared using "multiple *t*-tests," the probability of making a type I error (rejecting a true H_0) is increased as the number of comparisons made using independent *t*-tests increases.[4] The increase in α level can be calculated as follows:

Step 1
Number of comparisons:
X = no. of groups in experiment
C = no. of comparisons = $\frac{X(X - 1)}{2}$

Step 2
Corrected α level:
α corrected $= 1 - (1 - \alpha)^c$

Example: As shown above, with four groups, there can be a maximum of $4(4 - 1)/2 = 6$ paired comparisons. If the original α level was $P = .05$, the corrected α will be $1 - (1 - .05)^6 = .26$. Thus, there is now a .26 chance of inappropriately rejecting the null hypothesis (type I error) in at least one of the six comparisons made.[4] In most studies, this would be unacceptable! Should multiple *t*-tests be made among dependent groups, the corrected α levels are even greater than those calculated for independent groups.[4] Thus, multiple *t*-tests should not be accepted as a legitimate means of data analysis for the comparison of more than two groups.[4,6]

ANALYSIS OF VARIANCE

Analysis of variance (ANOVA) has long been an accepted method of comparing three or more groups from one experiment. The advantages of ANOVA over multiple *t*-tests include the following:[6] 1) The α level is held constant at the preset level with ANOVA, while the α level for multiple *t*-tests increases as the number of comparisons increases;[4] 2) one ANOVA is less cumbersome to calculate than are several *t*-tests; and 3) ANOVA is a more powerful data analysis method than is the *t*-test. ANOVA is the appropriate statistical method to test for differences among more than two groups. Often, it is assumed that ANOVA is used to determine if there is a difference among the means of these groups rather than among the groups' collective values. This is an incorrect assumption. While the mean describes a group in a meaningful way, it is simply a descriptor of the group. Many statistical references will discuss ANOVA as a comparison between means, but intragroup and intergroup variability is what is actually being analyzed.

It is also of value to understand how ANOVA relates to the theory of hypothesis testing. Without the tedium of a guided tour through the calculation of ANOVA, a simple explanation of ANOVA follows.

A test of the null hypothesis can be made in terms of two sets of differences (subjects participate in only one treatment, ie, subjects are "nested" within treatments). "One of these sets of differences is obtained by comparison of differences among treatment groups, referred to as external or *between-group* differences. The other set is obtained by comparison of differences among subjects receiving the same treatment within a treatment group, termed internal or *within-group* differences. Between-group differences are a result of the combined influence of the experimental treatment plus experimental error. Within-group differences are the result of experimental error alone."[7] The comparison ratio:

$$\frac{\text{Between-group differences}}{\text{Within-group differences}}$$

is sensitive to the effects of experimental treatment and can be written as:

$$\frac{\text{Treatment effect + experimental error}}{\text{Experimental error}}$$

Assuming that the experimental error rate estimates are approximately equal, any influence of treatment will result in a ratio that is greater than 1.[7] The above example of hypothesis testing illustrates the general theory behind the mathematical calculations of ANOVA.

Just as the *t*-test involves calculation of a *t*-statistic, which is compared with a critical *t*, ANOVA involves calculation of an *F*-ratio, which is compared with a critical *F*-ratio. The *F*-ratio answers the question, Is "the variability between the groups large enough in comparison to the variability of data within each group to justify the conclusion that two or more of the groups differ?"[6] If the variability between groups is large enough, we can conclude that there is a significant difference between groups. The *F*-ratio is defined as follows:

$$F\text{-ratio} = \frac{\text{Between-groups variance}}{\text{Within-groups variance}}$$

ANOVA is not just one simply defined computation. The experimental design possibilities are numerous with ANOVA. By using one test, several factors (eg, drugs, dose levels, dose times) can be analyzed for relationship at one time. The number of *F*-ratios calculated in an ANOVA is directly related to the number of factors in the experimental design. Thus, each ANOVA computation is unique to the experimental design being tested. It is the researcher's responsibility to ensure that the appropriate ANOVA is used, given the design of the study.

The assumptions for ANOVA are the same as those for the *t*-test.[4-6] To reiterate: 1) The populations from which the samples are drawn should approach normal distribution; 2) the variances of the populations from which the samples were drawn should be equal or nearly equal; and 3) the observations *within* groups must be independent.

These assumptions can usually be met by random sampling and by use of a good measurement scale.[6] The more that the above assumptions for ANOVA are violated, the more likely a type I or type II error will be made.[6]

As with the *t*-test, ANOVA is robust enough to be an appropriate test if the above assumptions are not strictly met (excepting the assumption of independence, which must be met at all times).[4,5] When the compared groups have equal values of n, population variances need not be homogenous. Also, normality of the population distributions may be violated to a limited degree without consequence.[4-6] Finally, because ANOVA is calculated using a parameter (variance), it is considered to be a parametric statistical analysis method and its use should be limited to interval and ratio scale data.

Thus, there exist many similarities between the *t*-test and ANOVA. This can further be extended to the calculated *t* from the *t*-test and to the *F*-ratio from ANOVA. If an ANOVA was being used instead of the *t*-test to compare two groups, it would be found that $F = t^2$ for these data.[4,5]

MULTIPLE COMPARISON METHODS

Following a significant *F* test, the next logical step would be to ask, Which of the groups compared in the ANOVA are significantly different? This question can be answered by the use of multiple comparison procedures. "All are essentially based upon the *t*-test but include appropriate corrections for the fact that more than one comparison is being made."[1]

There exist numerous legitimate methods of multiple comparison, each looking for unplanned yet "interesting" differences in the experimental data, but operating under a different set of rules and assumptions.[5] The test that is selected for use should be the test that meets the needs of the researcher and the design of the study. But overall, it is important to remember that the reason for using ANOVA and a multiple comparison method is ultimately to control the experimentwise error rate (the type I error rate for all comparisons) while at the same time making several different comparisons.[7] The experimentwise error rate can be limited by reducing the number of comparisons made or reducing the error rate within each comparison. Because most researchers do not want such imposing conditions placed on their work, as would be the case by limiting the number of comparisons allowed, the only other way to control the experimentwise error rate is to control the type I error rate within each comparison; hence, the purpose behind multiple comparison techniques. However, it is important to note that in reducing the type I error rate in such a way, there will be an increase in the type II error rate. Thus, before progressing, the researcher must determine which is more detrimental to the work, making type I errors or making type II errors.[7]

A summary flow chart of the selection of multiple comparison tests is shown (Figure). Use of this figure will help guide the researcher to select the test most appropriate for the experimental design tested and research questions asked. This flow chart, developed by Hopkins and Chadbourn[8] and modified by Keppel,[7] was intended to show the similarities and differences between some of the various multiple comparison methods. It should not be used as a "fixed and rigid plan for analysis."[7] For the purposes of this article, this chart serves as a logical guide to aid the reader in the understanding of multiple comparison methods.

Before any multiple comparison test, an α level is determined. Next, an F test is performed. If a significant F-ratio is obtained, the process of data analysis continues to determine which groups differ statistically. The test of Least Significant Difference (LSD) is an option if the researcher wishes to control the comparisonwise error rate (*individual* type I error rates for *each* comparison)[1] and if a small number of comparisons, relative to the total number of comparisons possible, are to be made. However, if the experimentwise error rate (type I error rate for *all* comparisons) must be held constant, other methods of multiple comparison must be considered.

There are two ways to control the experimentwise error rate. These include the layer or stepwise method and the experimentwise method. The layer method gradually adjusts the type I error rate. The experimentwise method holds the type I error rate constant for a set of comparisons. The Newman-Keuls test and Duncan test are examples of layer methods.

If an experimentwise method is selected, the type of comparisons to be made will determine the multiple comparison method selected. If comparisons are made only between a control group and experimental groups, the Dunnett test is an option of multiple comparison to consider. However, if the group comparisons are between any groups, there are other test options. The Dunn test could be considered if there are only a few comparisons to be made. If there are a large number of comparisons to be made, the Tukey test or the Scheffé test might be considered. The Tukey test assumes that the groups being compared are of equal size and is appropriate in the simple comparison of one group with another. The Scheffé test is based on the F statistic and thus is less affected by violations of the assumptions of normality and homogeneity of variances. Should comparisons be desired between complex combinations of groups, the Scheffé test will be sensitive in detecting real differences.[7]

While not included in the flow chart, the Bonferroni t-test is a multiple comparison method frequently used in medical literature. The Bonferroni t-test adjusts the preset α level by the number of comparisons to be made.[1,9]

$$\alpha_{adj} = \frac{\alpha_p}{n}$$

where p is the preset α level and n is the number of comparisons to be made. "If each comparison is made using the critical t corresponding to α_p/n, the error rate for all comparisons taken as a group will be at most α_p."[1] Thus the preset α level is protected. However, the Bonferroni t-test becomes very conservative as the number of comparisons made increases.[1]

Finally, as previously noted, confidence intervals may be more useful than multiple comparison tests in analysis of intergroup similarity.[2,3,9] "Confidence intervals: 1) Show the degree of uncertainty in each comparison in an easily interpretable way; 2) make it easier to assess the practical significance of a difference as well as the statistical significance; and 3) are less likely to lead non-statisticians to the invalid conclusion that nonsignificantly different sample means imply equal population means."[9]

The above discussion of multiple comparison methods and their uses is a basic overview of just a few of the possible options available to the researcher. There are other legitimate methods that have not been included in this discussion because of space limitations. Furthermore, statistical procedures and opinions on multiple comparison theory are continually evolving. The researcher is free to select whatever multiple comparison method is desired as long as the method is appropriate for the experimental design and research questions asked.

SUMMARY

In conclusion, when selecting the method for hypothesis testing, simplicity and familiarity must be pushed aside for assurance that the data being analyzed meet the defined assumptions required for use of a given test. For the t-test and ANOVA, these assumptions include normality of the populations from which the data come, homogeneity of the variances of the sample populations, and independence of the data points within a sample group.

If the experimental design consists of only two groups, the t-test is appropriate to test for a significant difference between these groups. However, if there are three or more groups to compare, the t-test is inappropriate because the preset level will increase with the number of comparisons made.

ANOVA is a powerful statistical test to determine simultaneously if there is a significant difference among three or more groups. While the F-ratio will tell if significance among any of the groups exists, it gives no information regarding which of the groups differs.

Thus, following a significant F-ratio, a multiple comparison test can be selected that will define which of the three or more groups is different. The multiple comparison method selection is based on the experimental design and the research questions asked.

REFERENCES

1. Glantz SA: *Primer of Biostatistics*, ed 2. New York, McGraw-Hill Book Co, 1987.

2. Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 3, Sensitivity, specificity, predictive value and hypothesis testing. *Ann Emerg Med* 1990;19:591-597.

3. Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 2, Descriptive analysis. *Ann Emerg Med* 1990;19:309-315.

4. Hopkins KD, Glantz GV: *Basic Statistics for the Behavioral Sciences*. Englewood Cliffs, New

Jersey, Prentice-Hall, Inc, 1978.

5. Sokal RR, Rolph FJ: *Biometry*, ed 2. New York, WH Freeman and Co, 1981.

6. Elston RC, Johnson WD: *Essentials of Biostatistics*. Philadelphia, FA Davis Co, 1987.

7. Keppel G: *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc, 1973.

8. Hopkins KD, Chadbourn RA: A schema for proper utilization of multiple comparisons in research and a case study. *Amer Educ Res J* 1967;4:407-412.

9. SAS Institute Inc: SAS/STAT™ User's Guide, Release 6.03 edition. Cary, North Carolina, SAS Institute Inc, 1988, p 1028.

Introduction to Biostatistics: Part 5, Statistical Inference Techniques for Hypothesis Testing With Nonparametric Data

Gary M Gaddis, MD, PhD
Monica L Gaddis, PhD†
Kansas City, Missouri

From the Departments of Emergency Health Services* and Surgery,† Truman Medical Center, University of Missouri-Kansas City School of Medicine.

Received for publication September 1, 1989. Accepted for publication May 28, 1990.

Address for reprints: Gary M Gaddis, MD, PhD, Department of Emergency Health Services, Truman Medical Center, 2301 Holmes, Kansas City, Missouri 64108.

Specific statistical tests are used when the null hypothesis (H_0) is to be tested using nonparametric nominal or ordinal data. With nominal data, experimental results are expressed by proportions or frequencies. Chi-square or related tests (the Fisher's exact test or the rows by columns test) are appropriate for testing H_0 with nominal data. Ordinal data permit arrangement of statistical results by rank. Rank-order tests used to test H_0 with ordinal data include the Mann-Whitney U, Kolmogorov-Smirnov, Wilcoxon, Kruskal-Wallis, and Friedman tests. The Kruskal-Wallis and Friedman tests permit multiple intergroup comparisons. Other rank-order tests permit only single intergroup comparisons. Specific details to guide the researcher in the proper selection of these tests are presented. [Gaddis GM, Gaddis ML: Introduction to biostatistics: Part 5, statistical inference techniques for hypothesis testing with nonparametric data. Ann Emerg Med September 1990;19:1054-1059.]

INTRODUCTION

Although the Student *t*-test and analysis of variance (ANOVA) are among the most powerful inferential statistical methods, they are not appropriate for nominal and ordinal data.[1]

Fortunately, statisticians have devised inferential statistical techniques appropriate for testing the null hypothesis (H_0) with nominal or ordinal data. The chi-square (χ^2) test, Fisher's exact test, and rows by columns test (R × C test) can be used when nominal data are to be analyzed. The Mann-Whitney *U*, Kolmogorov-Smirnov, and Wilcoxon rank tests can be used with ordinal data, and, like the *t*-test, are appropriate for single comparisons between groups. The Kruskal-Wallis and Friedman rank tests for ordinal data, like ANOVA, are appropriate for multiple comparisons between groups.

A conceptual framework for understanding the applicability of these nonparametric tests is provided in lieu of detailed examples, in the interest of brevity.

TESTS FOR PROPORTIONS AND FREQUENCIES OF NOMINAL DATA MATRICES

Chi-Square Tests

Chi-square tests are used to answer questions about rates, proportions, or frequencies.[2] In other words, χ^2 tests are used to tell whether a difference between populations or groups exists for the rate at which different outcomes occur. Chi-square tests are suited for the analysis of nominal data.

Two types of χ^2 tests exist. The χ^2 test of independence, also known as the χ^2 test of association, is commonly used in biomedical statistics and is used for comparison of two or more groups.[3] The χ^2 "goodness of fit" test is used to compare sample group data with data from a known population. This permits researchers to assess whether the sample group is drawn from the same population as the "standard of comparison," the known population.[4]

Chi-Square Test of Association

To perform a χ^2 test of association, data are arranged into a matrix, just

FIGURE 1. *Chi-square testing.*

	Outcome 1	Outcome 2	
Population 1	Cell A	Cell B	Row 1 Total
Population 2	Cell C	Cell D	Row 2 Total
	Column 1 Total	Column 2 Total	Grand Total = N

1. Obtain observed cell frequencies for each cell from the experimental data.
2. Calculate expected values for all cell frequencies:
   Expected value = row total x column total / grand total
3. For each cell, calculate value of:
   Observed frequency – Expected frequency
4. Square this value, and divide by expected frequency for each cell.
   Observed frequency – Expected frequency)^2 / Expected frequency
5. Sum these values for each cell to obain value of the χ^2 statistic.
6. For a 2x2 table only, use a shortcut formula:
   $\chi^2 = (AD-BC)^2 N / ([A+B] [C+D] [A+C] [B+D]$
   where N = grand total;
   A = observed value in cell A;
   B = observed value in cell B;
   C = observed value in cell C; and
   D = observed value in cell D.

This shortcut formula does *not* include the Yates correction factor.

1

as is the case for calculating sensitivity, specificity, and predictive value. Separate rows are made for each population being compared in the matrix, with separate columns for each type of outcome in the matrix. Unlike the case for sensitivity, specificity, and predictive value, the matrix can be not only 2×2 (Figure 1), but also any larger size. The row and column totals are derived, and then the expected frequencies for each cell (box) in the matrix are calculated (Figure 1). The individual values for the quantities (observed frequencies minus the expected frequencies) are calculated, then squared, and divided by the expected frequency in each cell. These values, calculated for each cell, are then summed (Figure 1). This yields a value of the χ^2 statistic, which is compared against a critical value table to determine whether statistical significance is achieved.

Chi-square tables list critical values for various degrees of freedom (*df*). The *df* in a χ^2 matrix is calculated as:

$$df = (\text{no. of rows} - 1)(\text{no. of columns} - 1)$$

The reason that this equation defines the number of degrees of freedom is as follows: when the row and column totals are known, after the cell frequencies for *df* cells are assigned, the remaining cell frequencies are constrained by the values of the previously assigned cell frequencies and the row and column totals.[4]

Chi-square critical value tables show that as the *df* increases, the critical value of the χ^2 statistic increases for any level of significance. This is a consequence of the nature of the χ^2 distribution for varying degrees of freedom.[4]

Six assumptions of the χ^2 test of association must be met if the test is to be properly applied.[2]

Only frequency data may be analyzed. (These data may be derived from data that are ordinal, interval, or ratio and transformed to nominal frequency data.)

Events must be independent within a sample group.

No cell of a 2×2 matrix, or less than 20% of the cells of larger matrices, may have a frequency of less than 5. (For 2×2 matrices, if a cell size is less than 5, Fisher's exact test should be used.)

There must be a logical or empirical basis for data classification into nominal groups.

The sum of expected frequencies for all cells must equal the sum of observed frequencies for all cells.

The sum of all observed frequencies minus the sum of all expected frequencies must equal 0.

Several limitations exist for the use of the χ^2 test.[2,3,5] The test cannot be used for paired or related samples because the assumption of independence is violated. Chi-square does not indicate the strength of an association. It merely indicates whether an association exists. For instance, $P < .001$ does not indicate a stronger association than $P < .05$.

When large frequency tables exist, the R×C test of independence using the G statistic can be substituted for the χ^2 test, for convenience. (The R×C test is discussed in more detail later in this paper.)

In matrices larger than 2×2, if any cell has an expected value of less than 1, or if more than 20% of the cells have an expected value of less than 5, no alternative exists except to combine categories or to collect more data. If categories are to be combined, they must be combined for a rational reason and not simply for convenience.[5]

Just as repeated sampling of normally distributed interval or ratio data yields differing estimates of the population mean and standard deviation, so repeated sampling of nominal data yields differing frequencies of each cell in a χ^2 matrix.[5]

The power of the χ^2 test increases as the number of individuals in the sample increases.[4]

Yates Correction for Continuity

The distribution of the χ^2 statistic is continuous, just as a temperature scale, yet χ^2 is applied to qualitative nominal data, which does not have a continuous distribution. This becomes a source of potential bias toward making a type I error in 2×2 matrices. Therefore, a correction factor has been derived by Yates to correct for this potential source of bias.[4,6] This correction involves subtracting 0.5 from the absolute value of the difference between each ob-

FIGURE 2. *Fisher's exact test.*

	Outcome 1	Outcome 2	
Population 1	Cell A	Cell B	Row 1 Total
Population 2	Cell C	Cell D	Row 2 Total
	Column 1 Total	Column 2 Total	Grand Total = N

1. Calculate the probability of the data matrix:
   $P = ([A+B]!)\ ([C+D]!)\ ([A+C]!)\ ([B+D]!)\ /\ (N!\ A!\ B!\ C!\ D!)$
2. Determine the distribution of the next less likely (or more extreme) matrix by subtracting 1 from the smallest cell frequency in the matrix. Then, fill in new values for the other three cell frequencies in the matrix by using new cell frequencies that permit the row and column totals to remain constant.
3. Calculate the probability of the revised data matrix with the same equation as above.
4. Repeat steps 2 and 3 until the value of the smallest cell equals 0.
5. Sum the individual *P* values obtained in the steps above. This yields the *P* value to be reported by the use of Fisher's exact test, in the case of a one-tailed test. For two-tailed tests, it is controversial whether it is proper to double the calculated Fisher *P* value because the distribution of probabilities for all possible outcomes is asymmetric.

2

FIGURE 3. *Mann-Whitney* U *test for groups with N > 20.*

1. Rank all N scores, from lowest to highest. Assign a rank of 1 to the lowest score and N to the highest score. Average the rank score for all ties within or between groups.
2. Sum the ranks of the smaller group to obtain R_s.
3. Calculate the value of *U*:
   $U = N_s \times N_1 + (N_s)\ (N_s + 1)\ /\ (2 - R_s)$
   where N_s = number of observations in the smaller size group;
   N_1 = number of observations in the larger group; and
   R_s = sum of ranks of the smaller group.
4. Calculate $U' = N_s \times N_1 - U$
5. Determine which is smaller, *U* or *U'*, and use the smaller value as the value of *U* for subsequent calculations.
6. When N is more than 20, the distribution of *U* approaches normality. Therefore, calculate:
   Mean $\mu_u = N_s \times N_1\ /\ 2$
   Standard deviation $\sigma_u = (N_s \times N_1)\ (N_s + N_1 + 1)\ /\ 12$
7. Calculate Z, which is compared with a table of the Z distribution to assess for statistical significance:
   $Z = U - \mu_u\ /\ \sigma_u$

3

served and expected cell frequency, yielding the formula $\chi^2 = (|[\text{Observed frequency} - \text{Expected frequency}]| - 0.5)^2\ /\ \text{Expected frequency}$.

The Yates correction is not needed when the matrix is larger than 2×2; the χ^2 statistic for the matrix does not reach statistical significance; or the number of individuals in the data matrix exceeds 40.[3,4,7]

χ^2 Goodness of Fit Tests

Chi-square goodness of fit tests are used to compare sample group data to known population frequency data. This lets researchers decide whether that sample group was drawn from the same population as the "standard of comparison," the known population.[4] A classic example of the use of this test involved testing whether the frequency of the ABO blood groups was the same in the general population as in a population of 223 sensitized Rh-negative women.[8] It was found that women of blood group A were over-represented in the population of sensitized women, when compared with the distribution of ABO blood groups in the general population. Therefore, it was concluded that having type A blood makes Rh sensitization in Rh-negative women more likely.

To calculate the χ^2 statistic for the goodness of fit test, data are arranged into a matrix. One column lists sample group cell frequencies. The other column lists population data cell frequencies. Thereafter, the mechanics of calculating the χ^2 statistic and determining significance from statistical tables are similar to the χ^2 test of independence, except that df = (number of rows of data − 1).

The χ^2 goodness of fit test is a one-column test, with population instead of sample data used to calculate expected values for each cell in the matrix. The assumptions and limitations of the χ^2 goodness of fit test are the same as for the χ^2 test of independence.

Fisher's Exact Test

Fisher's exact test is a variant of the χ^2 test of independence. It is used when the number of individuals in one cell of a 2×2 matrix of data from 20 to 40 individuals is less than five.[5-7] The assumptions of the test are the same as those of the χ^2 test.

The calculation of Fisher's exact test involves direct calculation of the probability value *P* for a one-tailed test (Figure 2).[6,7] This makes consideration of degrees of freedom for the data matrix irrelevant. Because the distribution of probabilities of data matrices is not symmetric, it is controversial whether it is appropriate to

Type of Data	N	Other Comments	Proper Test
Nominal, or other that has been transformed to nominal	> 20	2 x 2 or larger matrix, no cell frequency less than 5 Comparison of two or more experimental groups or types of outcomes	χ^2 test of independence
Same as above	> 20	2 x 2 matrix or larger, no cell frequency less than 5 Comparison of one experimental group to a previously defined population	χ^2 goodness of fit test
Same as above	20 or less	2 x 2 matrix only	Fisher's exact test
Same as above	> 20	2 x 2 matrix only; cell frequency less than 5 for one of the cells*	Fisher's exact test
Same as above	"Large"	For "large" data matrices or samples when logarithmic transformation facilitates calculation of results	RxC

*If the cell frequency is less than 5 for two or more of the cells, then it is controversial whether Fisher's exact test should be calculated.

4

FIGURE 4. *Selection of tests for nominal nonparametric data.*

double the calculated P value when a test is two-tailed.

Rows by Columns Test

The R×C test is a variant of the χ^2 test suited to very large sample sizes or very large data matrices.[6] The assumptions and limitations of this test are the same as for the χ^2 test. Calculation of the R×C test G statistic involves logarithmic transformation of the data, which is useful for handling calculations with large cell sizes. This test is often easier to use for slide rule or pocket calculator estimation of P values than the χ^2 test of independence when cell sizes are large. However, with the advent of widely available microcomputers, it is being used less frequently.

NONPARAMETRIC RANK TESTS FOR ORDINAL DATA

Rank tests are used for the analysis of ordinal data or for interval or ratio data that do not meet the assumptions of the t-test or ANOVA. Parametric statistical methods cannot be properly applied to nonparametric data, for which mean and standard deviation are improper and misleading.[1] The power of these nonparametric tests approaches 95% of the power of the t-test and ANOVA when applied to the same data.[7]

Rather than using the values of observations themselves, rank tests involve ranking parametric data from lowest to highest, then calculating an appropriate statistic. This number is then compared with a statistical table to assess possible statistical significance.[3] No assumptions are made about the nature of the data distribution in the population sampled. These tests are therefore known as "distribution-free" tests.[3] The Mann-Whitney U, Kolmogorov-Smirnov, Wilcoxon, Kruskal-Wallis, and Friedman tests are examples of nonparametric rank tests.

Mann-Whitney *U* Test

The Mann-Whitney U test was designed for analysis of ordinal data that are derived from independent samples, exist at discrete levels of magnitude in the distribution, and are not grouped into a cumulative frequency distribution. In other words, individual scores or data values are not lumped together into groups for further analysis. The Mann-Whitney U test is a nonparametric analog of the t-test. A single comparison is made between two nonpaired groups (which need not be of equal size). The test can be performed as one-tailed or two-tailed, as the hypothesis being tested warrants.[3,5,7]

The test examines whether the distribution of ranked responses between the two samples being compared are significantly different. It does this by the calculation of a U statistic that is based on the rank order of all data points. From this U value, a probability value is derived.

The calculation of U varies by the size of the sample studied and can be referenced[7] (Figure 3). The presence of ties of rank *between* members of *different* groups affects the distribution of U only slightly but can be corrected for. However, the effect of ties between members of different groups is so small that this correction is usually omitted.[7]

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is used to see whether two independent samples of data have been drawn from the same or different populations. It is a nonparametric analog of the nonpaired t-test. However, the Kolmogorov-Smirnov test is applied to data grouped into cumulative distribution ranges.[7] An example of a cumulative distribution grouping familiar to many emergency physicians is the manner in which the Glasgow Coma Score is used to help determine the Trauma Score.[9] Point values are assigned to various ranges of Glasgow Coma Scores and combined with scores from the respiratory rate, systolic blood pressure, capillary refill score, and respiratory expansion quality toward the calcu-

TABLE. *Selection of tests for ordinal nonparametric data*

Type of Data	No. of Groups Compared	No. of Comparisons	Do Subjects Participate in all Treatment Groups? (Do Subjects "Serve as Their Own Controls"?)	Test
Ordinal; or interval or ratio data that do not meet assumptions of the *t*-test, data not grouped into cumulative frequency distribution	2	1	No	Mann-Whitney *U*
Ordinal; or interval or ratio data as above, with data arranged as a cumulative frequency grouping	2	1	No	Kolmogorov-Smirnov
Ordinal; or interval or ratio data not meeting *t*-test assumptions	2	1	Yes	Wilcoxon
Ordinal; or interval or ratio data not meeting ANOVA assumptions	3 or more	2 or more	No	Kruskal-Wallis
Ordinal; or interval or ratio data not meeting ANOVA assumptions	3 or more	2 or more	Yes	Friedman

lation of the Trauma Score.

When a sufficiently large difference exists between the ranking of the two sample cumulative distributions under comparison, H_0 is rejected. The calculation of the Kolmogorov-Smirnov test is detailed by Siegel[7] and is conceptually similar to the calculation of the Mann-Whitney *U*.

Wilcoxon Test

The Wilcoxon test is a nonparametric rank test analogous to a paired *t*-test because sample sizes in both groups compared are equal, and all individuals are represented in both groups being compared.[3] Therefore, samples are not independent, they are related. The consequence of this nonindependence is that tests such as the Mann-Whitney *U* are improper. In addition to being suitable for analysis of nonindependent ordinal data, the Wilcoxon test can be used to analyze interval or ratio data that do not meet the normality or homogeneity of variances assumptions of the paired *t*-test.[1,5]

As is true for the other rank order tests, the size of the data samples affects the mechanism of calculating the test statistic W for the Wilcoxon test. When 20 or more subjects are studied, the distribution of the W statistic approaches normality, just as the distribution of the Mann-Whitney *U* statistic approaches normality when N exceeds 20.[3] The calculation of the Wilcoxon test is easily referenced and has conceptual and methodologic similarities to the calculations of the Mann-Whitney *U* test.

Kruskal-Wallis Test

The Kruskal-Wallis test is the nonparametric analog of a one-way ANOVA, because it can be used when three or more groups of subjects are compared and when subjects are not permitted to participate in more than one group.[3] This test has its own unique formula for the calculation of the Kruskal-Wallis statistic H. The value of H is compared with the χ^2 distribution, with *df* = (number of treatment groups − 1) if the sample size is not small. When H exceeds the critical value of the χ^2 statistic for the appropriate number of degrees of freedom, H_0 is rejected and groups compared are deemed statistically different.[3]

Friedman Test

The Friedman test is a rank test used to analyze data obtained when subjects participate in three or more treatment groups. Observations are ranked, and the Friedman test statistic X^2r is calculated in a manner similar to the Kruskal-Wallis H. The value of X^2r compared with the χ^2 critical value table to assess possible statistical significance, just as occurs with the Kruskal-Wallis test. The Friedman test is analogous to repeated measures ANOVA because subjects participate in more than one treatment group.[3]

SUMMARY

The proper use of various inferential statistical tests for nonparametric data has been discussed. Figure 4 and the Table summarize key information to guide the reader in the selection of the proper test technique.

REFERENCES

1. Gaddis ML, Gaddis GM: Introduction to biostatistics: Part 4, Statistical inference techniques in hypothesis testing. *Ann Emerg Med* 1990;19:820-825.

2. Isaac S, Michaels WB: *Handbook in Research and Evaluation*. San Diego, Edits Publishers, 1979.

3. Glantz SA: *Primer of Biostatistics*, ed 2. New York, McGraw-Hill Book Co, 1987.

4. Bahn AK: *Basic Medical Statistics*. New York, Grune and Stratton, 1972.

5. Elenbaas RM, Elenbaas JK, Cuddy PG: Evaluating the medical literature, part II: Statistical analysis. *Ann Emerg Med* 1983;12:610-620.

6. Sokal RR, Rohlf FJ: *Biometry.* ed 2. New York, WH Freeman, 1981.

7. Siegel S: *Nonparametric Statistics for the Behavioral Sciences.* New York, McGraw-Hill, 1956.

8. Lucia SP, Hunt ML, Talbot JC: On the relationship of blood group A to Rh immunization and the occurrence of hemolytic disease of the newborn. *Science* 1949;110:309-330.

9. Champion HR, Sacco WJ, Carnazzo AJ, et al: Trauma score. *Crit Care Med* 1981;9:672-676.

ERRATUM

Figure 3 in the article, "Introduction to Biostatistics: Part 5, Statistical Inference Techniques for Hypothesis Testing With Nonparametric Data" [September 1990;19:1054-1059], contains two errors.

The equation $U = N_S \times N_1 + (N_S)(N_S + 1) / (2 - R_S)$ should read

$$U = (N_S \times N_1) + \left[\frac{(N_S)(N_S + 1)}{2}\right] - R_S$$

The equation $\sigma_u = (N_S \times N_1)(N_S + N_1 + 1) / 12$ should read

$$\sigma_u = \sqrt{(N_S \times N_1)(N_S + N_1 + 1) / 12}$$

SPECIAL CONTRIBUTION
biostatistics

Introduction to Biostatistics: Part 6, Correlation and Regression

Correlation and regression analysis are applied to data to define and quantify the relationship between two variables. Correlation analysis is used to estimate the strength of a relationship between two variables. The correlation coefficient r *is a dimensionless number ranging from* -1 *to* $+1$. *A value of* -1 *signifies a perfect negative, or indirect (inverse) relationship. A value of* $+1$ *signifies a perfect positive, or direct relationship. The* r *can be calculated as the Pearson-product* r, *using normally distributed interval or ratio data, or as the Spearman rank* r, *using non-normally distributed data that are not interval or ratio in nature. Linear regression analysis results in the formation of an equation of a line* ($Y = mX + b$), *which mathematically describes the line of best fit for a data relationship between X and Y variables. This equation can then be used to predict additional dependent variable values* ($\hat{Y}$), *based on the value or the independent variable X, the slope m, and the Y-intercept b. Interpretation of the correlation coefficient* r *involves use of* r^2, *which implies the degree of variability of Y due to X. Tests of significance for linear regression are similar conceptually to significance testing using analysis of variance. Multiple correlation and regression, more complex analytical methods that define relationships between three or more variables, are not covered in this article. Closing comments for this final installment of this introduction to biostatistics series are presented. [Gaddis ML, Gaddis GM: Introduction to biostatistics: Part 6, correlation and regression. Ann Emerg Med December 1990;19:1462-1468.]*

Monica L Gaddis, PhD*
Gary M Gaddis, MD, PhD†
Kansas City, Missouri

From the Departments of Surgery* and Emergency Health Services,† University of Missouri-Kansas City School of Medicine, Truman Medical Center, Kansas City, Missouri.

Received for publication August 23, 1990. Accepted for publication September 10, 1990.

Address for reprints: Monica L Gaddis, PhD, Department of Surgery, University of Missouri-Kansas City School of Medicine, Truman Medical Center, 2301 Holmes, Kansas City, Missouri 64108.

INTRODUCTION

In research or clinical practice, the physician is often asked to relate two or more variables to predict an outcome. This is exemplified by how risk factors are devised for prediction of chance of disease occurrence. Risk factors associate physical or behavioral characteristics with illness in a manner that is easily understood by the lay public. An example is the association between cigarette smoking, hypertension, and heart disease.

The statistical methods used to define or describe such risk factor-disease relationships are termed correlation and regression analysis. While the terms correlation and regression are often used together, they represent separate steps in the process of relationship analysis. Correlation analysis provides a quantitative way of measuring the strength of a relationship between two variables. Regression analysis is used to mathematically describe that relationship, with the ultimate goal being the development of an equation for prediction of one variable from one or more other variables.

This final installment of the Introduction to Biostatistics series will provide an independent discussion of correlation and regression analysis. Final comments regarding the entire biostatistics series also are included.

CORRELATION ANALYSIS

In nature, one can often see that two or more variables are related or correlated. The latitude of a city and its average daily temperature, intelligence quotient of a student and his grade point average, drug dosage administered and the resultant physical response, and caloric consumption and weight gain or loss are examples. While some relationships between two or more variables appear obvious, they by no means serve as perfect

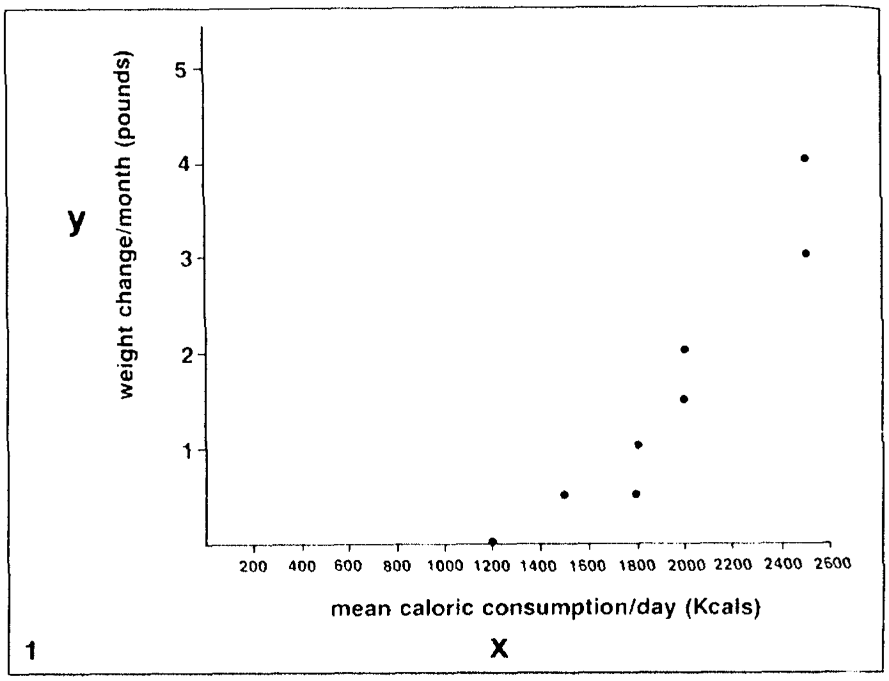


FIGURE 1. *Scatter diagram for sample data given in Table 1 (caloric consumption vs weight change).*

TABLE 1. *Sample data: Caloric consumption versus weight change*

Patient	(X) Mean Caloric Consumption/Day	(Y) Weight Change/ Month
1	1,200	0.0
2	1,500	0.5
3	1,800	0.5
4	2,000	1.5
5	2,500	4.0
6	1,800	1.0
7	2,500	3.0
8	2,000	2.0

predictors of outcome. The relationships may be confounded by extraneous variables that can adversely affect a "perfect" correlation, causing variability of the responses. For example, while it seems obvious that the closer to the equator a city lies, the higher its average daily temperature should be, altitude and weather patterns may influence or change the expected relationship. In addition, individual metabolic rate and physical activity will certainly affect weight gain or loss, regardless of controlled caloric consumption. Thus, while the direction of a relationship is often intuitive, the strength of that relationship is not.[1]

The discussion of the degree of relationship between two random variables is also a discussion of the correlation between those variables because "correlation is a relation."[2] However, it is very important to remember that correlation simply recognizes a relation. Correlation does not imply causation!

The degree of correlation between two variables is estimated in the following manner: first, visualization of the relationship is best obtained by placing the variables in question in graphic form. This graphic representation is termed a "scatter diagram."[1,3] A scatter diagram is made by use of an X (horizontal) and Y (vertical) axis graph. Figure 1 is a scatter diagram of mean caloric consumption per day versus weight change per month (Table 1). While the total number of patients is small, a relationship emerges from the scatter diagram that indicates that as caloric consumption increases, so also does weight. In Figure 2, some of the various relationships seen between two random variables in a scatter diagram are shown. These include a) positive (direct) linear; b) negative (indirect) linear; c) and d) curvilinear; e) no relationship; and f) exponential.[1,4,5] However, the scatter diagram is limited to simply describing the appearance or pattern of the relationship.

To quantify the relationship in a meaningful manner, a correlation coefficient, a numerical value that describes the strength of the relationship, must be calculated.[1-8] The correlation coefficient *r* is a dimensionless number ranging from −1 to +1, with −1 depicting a perfect negative linear relationship and +1 depicting a perfect positive linear relationship.[1-8] The closer that *r* is to 0, allegedly the weaker the relationship. However, "a small correlation between two variables can also be due to 1) little *linear* association between the two variables, or 2) large errors in the measurement of the variables."[6]

There are several methods available to calculate a correlation coefficient. Included in this discussion are the Pearson product-moment *r* and the Spearman rank *r*, both commonly used in clinical data analysis.

Pearson Product-Moment *r*

The Pearson *r* is used to quantify the strength of a linear relationship between two continuous variables (of interval or ratio scales) that are from normally distributed populations.[1,3,5]

Before calculation of the Pearson *r* is shown, it is important to understand the concept of covariance. Recall from part 3 of this series[9] that variance is a measure of the variability or dispersion of a single random variable.[6] Covariance, an extension of this, is defined as "a measure of how two random variables vary together."[6] Using the example in Table 1, computation of covariance is as

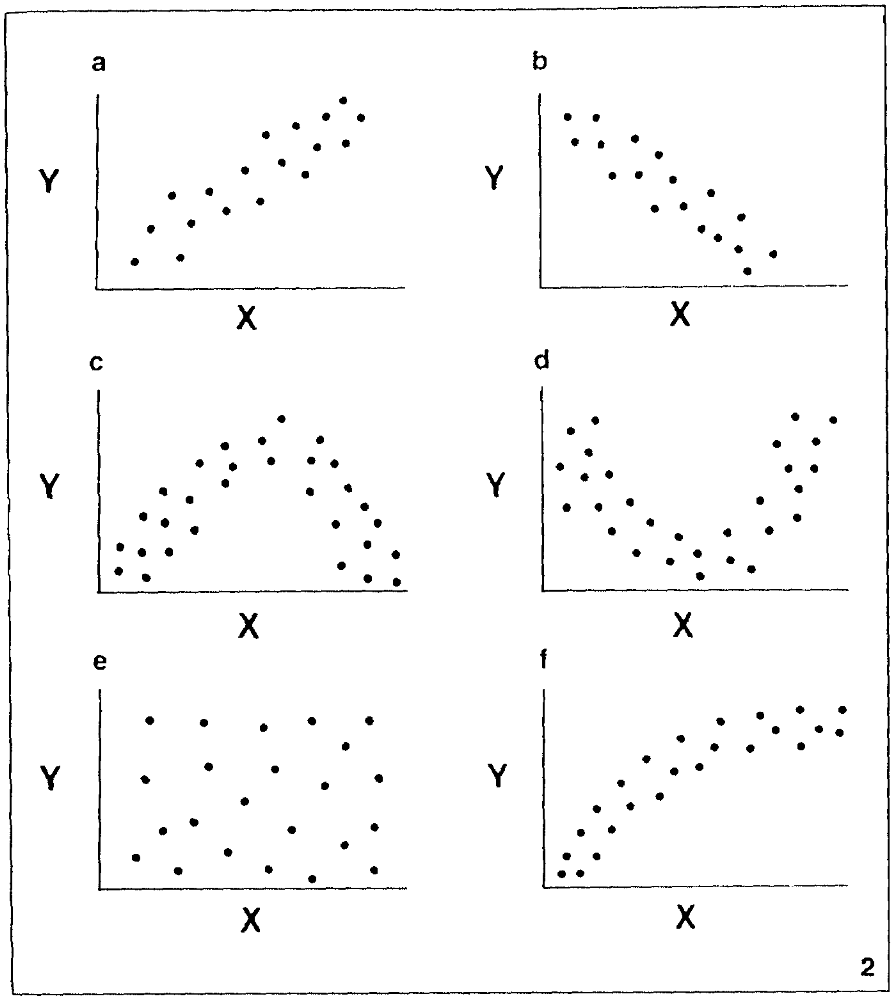


FIGURE 2. *Scatter diagram relationships: a) positive (direct) linear; b) negative (indirect) linear; c) curvilinear; d) curvilinear; e) no relationship; and f) exponential.*

follows:[1,3]

1. Calculate the mean values of the X and Y variables: ($\overline{X}$ and $\overline{Y}$).
2. Subtract the corresponding mean from each individual value: ($x_i - \overline{X}$) and $y_i - \overline{Y}$).
3. Calculate the product ($x_i - \overline{X}$) ($y_i - \overline{Y}$) for each variable pair.
4. Sum the products calculated in step 3.
5. Divide the sum of the products by the total number of variable pairs (n) minus 1: (n − 1).

Covariance = S_{xy} = $[\Sigma\ (x_i - \overline{X})\ (y_i - \overline{Y})]\ /\ (n - 1)$

A positive covariance implies that when one variable is greater than its mean, so too is the corresponding variable. A negative covariance implies that when one variable is greater than its mean, the corresponding variable will be less than its corresponding mean. Covariance relates to the equation for the Pearson *r* in that it is the numerator in the equation for *r*. The denominator is obtained by calculating the standard deviation (SD) for the values of variable x (S_x), the SD for the values of variable y (S_y), and finding their product ($S_x \circ S_y$) (Table 2). Therefore, the equation for the Pearson product moment *r* is:[1,3,5,7]

$$r = S_{xy}\ /\ S_x \circ S_y$$

The calculated Pearson *r* in the example (Table 2) is .94. This implies some degree of positive or direct linear correlation.

Spearman Rank Order *r*

When one or more of the variables being analyzed for strength or direction of a relationship or trend is not of an interval of ratio scale, is not drawn from normally distributed population, or does not possess a linear relationship, a nonparametric correlation technique must be used. One such method is the Spearman rank order correlation coefficient. The Spearman rank *r* is based on the rank order of the individual data points and not the actual numerical values. The Spearman rank *r* is calculated as follows (Table 3):[3,5]

1. The values of the two variables are ranked in ascending or descending order (for ties, the rank is the average rank).
2. Calculate the difference between the ranks for each pair of data.
3. Calculate the sum of the squared differences (from step 2) and multiply by 6.
4. Divide the number calculated in step 3 by $n(n^2 - 1)$, where n equals the number of pairs of data.
5. Subtract the number found in step 4 from 1.

Spearman rank

$$r = 1 - 6\ \Sigma\ (d^2)\ /\ n\ (n^2 - 1)$$

Note that Pearson *r* and Spearman rank *r* are similar for the same set of data. This will be true, especially for large data sets where the number of pairs of data is more than 50.[1]

Interpretation and Use of the Correlation Coefficient

As previously stated, while a strong correlation might exist between two independent variables, this does not imply that one event causes the other. The reasoning behind this is threefold;

1. Which comes first, the chicken or the egg ie, ... does X cause Y, or does Y cause X? The *r* calculation does not have the capabilities to determine the order of the causal relationship.
2. A variable other than X or Y may influence the relationship.
3. Complex relationships as found in the biomedical sciences can rarely be explained by a simple two variable relationship.[1]

Thus, *r* (whether from Pearson or Spearman) is simply a descriptor of the strength and direction of the relationship between the two variables

TABLE 2. *Calculation of Pearson product-moment correlation coefficient using sample data from Table 1*

Patient	x		$\bar{X}$		$(x_i - \bar{X})$	y		$\bar{Y}$		$(y_i - \bar{Y})$	$(x_i - \bar{X})(y_i - \bar{Y})$
1	1,200	−	1,912.5	=	−712.5	0.0	−	1.56	=	−1.56	1,111.5
2	1,500	−	1,912.5	=	−412.5	0.5	−	1.56	=	−1.06	437.25
3	1,800	−	1,912.5	=	−112.5	0.5	−	1.56	=	−1.06	119.25
4	2,000	−	1,912.5	=	87.5	1.5	−	1.56	=	−0.06	−5.25
5	2,500	−	1,912.5	=	587.5	4.0	−	1.56	=	−2.44	1,433.5
6	1,800	−	1,912.5	=	−112.5	1.0	−	1.56	=	−0.56	63.0
7	2,500	−	1,912.5	=	587.5	3.0	−	1.56	=	−1.44	846.0
8	2,000	−	1,912.5	=	87.5	2.0	−	1.56	=	−0.44	38.5
	X = 1,912.5					Y = 1.56					Σ 4,043.75

Step 1 $S_{xy} = \Sigma\,[\,(x_i - \bar{X})\,(y_i - \bar{Y})]\,/\,(n - 1) = \frac{4{,}043.75}{8 - 1} = 577.7$

Step 2 $S_x = \frac{\Sigma\,(x_i - X)^2}{(n - 1)} = \frac{1{,}408{,}750}{8 - 1} = 448.6$

Step 3 $S_y = \frac{\Sigma\,(y_i - Y)^2}{(n - 1)} = \frac{13.22}{8 - 1} = 1.37$

Step 4 $r = \frac{S_{xy}}{S_{x\circ}\,S_y} = \frac{577.7}{(448.6)\,(1.37)} = 0.94$

TABLE 3. *Calculation of Spearman rank order correlation coefficient using sample data from Table 1*

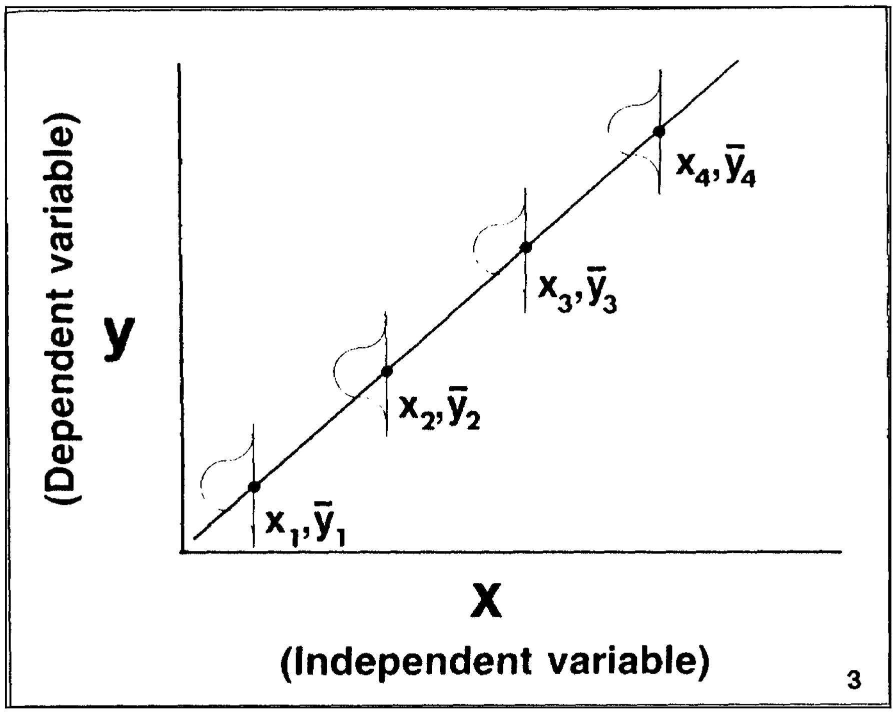
Patient	Rank X	Rank Y	d	d²
1	1.0	1.0	0	0
2	2.0	2.5	−0.5	0.25
3	3.5	2.5	1.0	1.00
4	5.5	5.0	0.5	0.25
5	7.5	8.0	−0.5	0.25
6	3.5	4.0	−0.5	0.25
7	7.5	7.0	−0.5	0.25
8	5.5	6.0	−0.5	0.25
				Σ(d²) = 2.5

$$r_s = 1 - \frac{[6\ (\Sigma(d^2))]}{n(n^2 - 1)} = 1 - \frac{[(6)(2.5)]}{(8)\ (63)} = 1 - 0.030 = 0.97$$

in question and nothing more. Just as in hypothesis testing, $P < .001$ does not imply greater significance than $P < .05$, so a Pearson or Spearman r of 0.99 does not imply a more likely causation than an r of 0.95. However, correlation analysis may guide the researcher in determining causation. Additional studies to aid in defining the causative variables may be developed. These studies may be based on knowledge gained from prior correlation analysis. While the value of r describes the strength of a relationship between two variables, an r of 0.5, for example, does not imply that "the strength of that relationship is 'halfway' between zero correlation (no relationship) and a perfect positive correlation (1.0)."[3] The correlation coefficient r can be squared and then used to estimate "the percent of variation in one variable that is explained by variation in the other variable."[3]

In the sample data given (Table 1), the calculated Pearson r is 0.94. The square of r is 0.88. This implies that 88% of the variability in weight gain or loss can be attributed to variability in the amount of calories consumed. Otherwise stated, the amount of calories consumed provides us with approximately 88% of the information needed to predict weight gain or loss. Finally, r is calculated using sample data. It is only an estimate of the true correlation coefficient, rho (ρ), between two population variables, just as the mean ($\bar{X}$) is a sample mean and is only an estimate of the true population mean mu (μ).

Should one wish to determine whether r is significant (ie, whether a significant correlation exists between two sample variables), hypothesis testing is indicated. This hypothesis testing answers the question, "Is r different from 0 only because of chance variation (sampling error), or because the true population correlation coefficient ρ is not 0?"[3] To complete the hypothesis test for r, the value of r and the degrees of freedom, (n − 2) (where n = the number of pairs of data correlated), are applied to the appropriate table of critical



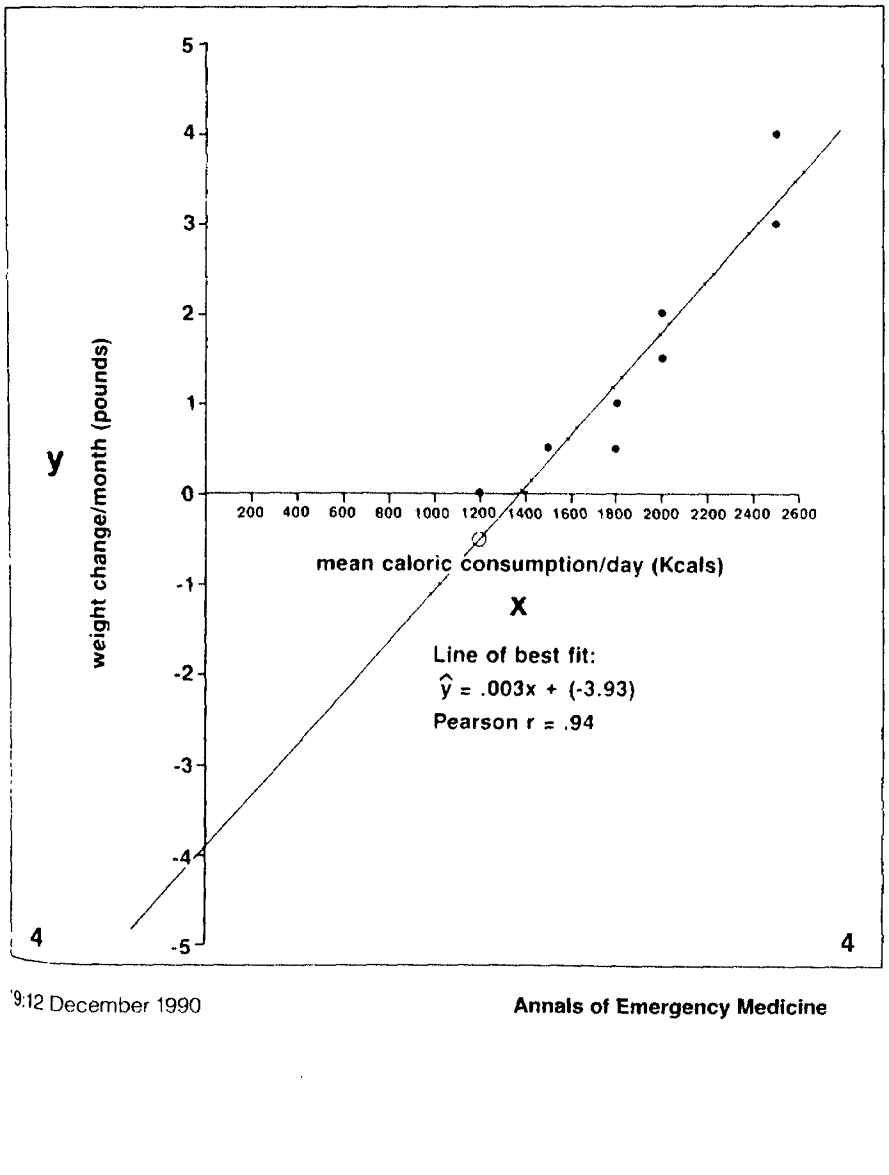


FIGURE 3. *For a given value of the independent variable (X), there will be a subpopulation of dependent variables (Y) that displays a normal distribution. The mean value of each subpopulation will form a point on the regression line.*

FIGURE 4. *Regression line for sample data given in Table 1 (caloric consumption vs weight change).*

values of *r*. If the calculated *r* is greater than r_{critical} from a statistical table at a predetermined significance level, *r* is determined to be significantly different from zero, and thus a significant correlation exists. Any correlation coefficients reported in research studies should include the corresponding level of significance.

REGRESSION ANALYSIS

Like correlation analysis, regression analysis is also used to describe the relationship between two or more variables. However, regression, in the process of mathematically describing a relationship, also develops a means of prediction of one variable based on the value of a second variable.[8] "In regression analysis, the relationship between two variables (or more) is expressed by fitting a line or curve to the pairs of data points" as seen in a scatter diagram.[3] The "predictor variable," or independent variable (X), is that which is selected and manipulated by the investigator. The "predicted" variable, or dependent variable (Y), is that which is influenced by X.

Simple linear regression is used to describe and predict a linear relationship between two variables, one independent and one dependent. (While regression analysis can be used in the analysis and prediction of nonlinear relationships as well as multiple variable data sets containing two or more independent variables, this is beyond the scope of this discussion.)

The assumptions underlying simple linear regression include:[4]

1. The values of the independent variable X are set by the investigator. In the process of data collection, the "pre-set" levels of X should not be changed.
2. The independent variable X should be measured without experimental error.
3. For each value of X, there is a sub-

population of Y variables that are normally distributed up and down the Y-axis (Figure 3).

4. The variances of the subpopulations of Y are homogeneous (Figure 3).
5. The means of the subpopulations of Y lie on a straight line (thus, the assumption of linearity is met) (Figure 3).
6. All values of Y are independent from one another, though they are dependent on X.

The Regression Line

The process of linear regression involves defining the "line of best fit," or the line that best defines the linear relationship of the data. This line is the one that passes through the center of the data points.[8] For linear regression, this line is defined by the equation for a straight line.

Recall that the equation for a straight line is:

$$Y = mX + b$$

where Y is the dependent variable, X is the independent variable, m is the slope of the line, and b is the y-intercept of the line.

Figure 4 shows the line of best fit for the data set described in Table 1. This line is also used as a predictor for additional data (dependent variables). When the line of best fit is used to predict Y, the equation is simply modified as $\hat{Y} = mX + b$. $\hat{Y}$ implies a predicted (or estimated) dependent variable, based on the defined slope, y-intercept, and independent variable values.

The calculation of the regression line (line of best fit) is performed by the "least-squares method." This is accomplished by calculating the equation for the line such that the sum of the squared deviations for each data point from the predicted line of best fit is as small as possible.[3,6] In other words, the least-squares method calculates the line that comes the closest to running through all of the data points. Therefore, the least-squares method derives the equation of the line that relates the linear relationship between all of the data points with a minimum of error. For details of the calculation of the regression line by the least-squares method, consult any of the referenced statistics texts.[1,3,5-8]

The prediction of a dependent variable by use of the regression line of best fit is also related to the correlation coefficient *r*. The letter *r* is used to designate the correlation coefficient, yet *r* implies regression. This is indicative of the relationship between correlation and regression. An *r* of +1 or −1 implies a perfect correlation between X and Y and also implies that predicted values of Y would assume a straight line. If *r* = 0, the best predicted value of Y is the mean of Y. Thus, the closer the absolute value of *r* is to 1.0, the closer the predicted values of Y will lie to the regression line.[2]

Interpretation of the Regression Line

Tests of significance for linear regression are conceptually similar to analysis of variance (ANOVA).[2] Recall from the prior discussion of ANOVA[9] that the total variance in ANOVA can be partitioned into the "between groups (experimental) variance," and the "within groups (error) variance." Alternatively expressed, total variance = experimental variance + error variance. Regression analysis is similar in that the regression variance is partitioned in the same way. Variance due to the regression process *per se* is analogous to ANOVA experimental variance. Variance due to the "residuals," or the deviations of the individual data points from the regression line, is analogous to ANOVA error variance. Therefore, for regression, total variance = regression variance + residual variance. The variance is then put into the form of the sums of squares (SS), so that for ANOVA, the equation is:[9]

$$SS_{\text{total}} = SS_{\text{between groups}} + SS_{\text{within groups}}$$

and for regression analysis, the equation is:[2]

$$SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{residual}}$$

Once the sums of squares are calculated, the F ratio can be derived. For ANOVA, the F ratio is calculated using SS values and the degrees of freedom (*df*) applicable:

$$F = [(SS_{\text{between groups}}) / df] / [(SS_{\text{within groups}}) / df]$$

and for regression, the F ratio is calculated:

$$F = [(SS_{\text{regression}}) / df] / [(SS_{\text{residual}}) / df]$$

In both cases, the $F_{\text{calculated}}$ is then compared with F_{critical}, found in a table of critical F ratios. For ANOVA, when $F_{\text{calculated}}$ is greater than F_{critical}, the difference between group means is concluded to be different from zero, and there exists a significant difference between at least two of the groups compared. For regression, when $F_{\text{calculated}}$ is greater than F_{critical}, the slope of the regression line is statistically different from zero. Calculation by this method is cumbersome and often tedious. Most statistical packages for computers that contain linear or multiple regression calculate both the regression line and the test of significance for that line.

CONCLUDING REMARKS

This installment concludes this six-part series of Introduction to Biostatistics articles. Objectives of the series have included the introduction of basic concepts of commonly used biomedical statistical tests to facilitate comprehension by persons with little or no background in biostatistics. We hope that this series has increased the understanding of biomedical statistics among clinicians, residents, and residency faculty.

Part 1 of this series in the January issue of *Annals* included introductory comments and discussed the frequency of errors of use of biomedical statistics in the medical literature. The concepts of sample and population were introduced. Types of data were discussed. As has been shown in subsequent installments of the series, it is crucial to understand what type of data is being analyzed in order to properly select both descriptive statistics and inferential statistical tests for significance testing.

Part 2 in March discussed descriptive statistics, including measures of central tendency and measures of variability. It was shown that certain descriptive statistics are inappropriate descriptors of certain types of data. For example, the mean value for an ordinal data sample is at best misleading and represents an erroneous use of the mean. Confidence intervals were also discussed.

Part 3 in May covered sensitivity, specificity, and predictive value of clinical tests. These commonly applied clinical principles were then used as an analogy to facilitate understanding of hypothesis testing. In this manner, the concepts of type I error (α), type II error (β), and statistical power (1-β) were introduced.

Part 4 in July presented inferential statistical techniques appropriate for parametric data, including the Student t-test and ANOVA. These techniques are appropriate only for interval or ratio data that meet the assumptions of these tests.

Part 5 in September continued discussion of inferential statistical techniques, covering tests properly applied to nonparametric data. Included were the χ^2 and Fisher's exact test for nominal data, and the "distribution free" rank tests such as the Mann-Whitney U, Wilcoxon, Kolmogorov-Smirnov, Kruskal-Wallis, and Friedman tests. These rank tests are appropriate for analysis of ordinal data and also for interval or ratio data that do not meet the assumptions of the t-test or ANOVA.

Finally, this article, part 6, has discussed correlation and regression, two data analysis techniques used to quantify and define the relationship between two variables. It was shown that tests for significance for linear regression are conceptually equivalent to ANOVA tests for significance of parametric data.

This series did not cover nonlinear and multiple correlation and regression, or multivariate ANOVA. Some specific tests were omitted from discussion in some sections in the interest of brevity. The intent of this series was to introduce basic concepts, not to provide a comprehensive review of all possible statistical tests.

The authors acknowledge the contributions of numerous individuals who inspired and aided the completion of this series. Dr Hal Morris of the Department of Health, Physical Education, and Recreation at Indiana University first brought the frequency and consequences of improper statistical analysis in biomedical sciences to our attention and instructed us in much of the information presented here. Dr Carl Rothe of the Department of Physiology and Biophysics of the Indiana University School of Medicine provided further practical teaching and detailed analysis methods. The Department of Emergency Medicine at the Wright State University School of Medicine provided us our first opportunity to contribute to the statistical education of emergency medicine residents and faculty. Dr Joseph Waeckerle recognized the importance of proper statistical analysis, provided a forum to present this information, and lent editorial assistance to the project. We also thank Jayna Ross and Kathy Wagner of the Department of Emergency Health Services at the University of Missouri-Kansas City School of Medicine for preparation of the manuscripts.

REFERENCES

1. Hopkins KD, Glass GV: *Basic Statistics for the Behavioral Sciences.* Englewood Cliffs, New Jersey, Prentice-Hall Inc, 1978.

2. Kerlinger FN, Pedhazur EJ: *Multiple Regression in Behavioral Research.* New York, Holt, Rinehart and Winston, Inc, 1973.

3. Knapp RG: *Basic Statistics for Nurses,* ed 2. New York, John Wiley and Sons, 1985.

4. Daniel WW: *Biostatistics: A Foundation for Analysis in the Health Sciences.* New York, John Wiley and Sons, 1974.

5. Glantz SA: *Primer of Biostatistics,* ed 2. New York, McGraw-Hill Book Co, 1987.

6. Elston RC, Johnson WD: *Essentials of Biostatistics.* Philadelphia, FA Davis, Co, 1987.

7. Sokol RR, Rohlf FS: *Biometry,* ed 2. New York, WH Freeman and Co, 1981.

8. Kviz FJ, Knapp KA: *Statistics for Nurses, An Introductory Text.* Boston, Little, Brown and Co, 1980.

9. Gaddis ML, Gaddis GM: Introduction to biostatistics: Part 4, statistical inference techniques in hypothesis testing. *Ann Emerg Med* 1990;19:820-825.